

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

From Data to Evaluation: Investigating the Limits of the Individual-Control Model in AI Governance

Permalink

<https://escholarship.org/uc/item/7135b09g>

ISBN

9798288862052

Author

Whitney, Cedric

Publication Date

2025-05-15

Peer reviewed|Thesis/dissertation

From Data to Evaluation: Investigating the Limits of the Individual-Control Model in AI
Governance

By

Cedric Deslandes Whitney

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Deirdre Mulligan, Chair

Professor Jenna Burrell

Associate Professor Niloufar Salehi

Spring 2025

From Data to Evaluation: Investigating the Limits of the Individual-Control Model in AI
Governance

Copyright 2025
by
Cedric Deslandes Whitney

Abstract

From Data to Evaluation: Investigating the Limits of the Individual-Control Model in AI Governance

by

Cedric Deslandes Whitney

Doctor of Philosophy in Information Science

University of California, Berkeley

Professor Deirdre Mulligan, Chair

Technical fixes for AI harm are usually framed through an *individual-control* lens: give people notice, let them opt out, erase their row, and the problem is solved. Building on relational theories of data governance, this dissertation probes the limits of that premise across the full machine-learning pipeline. It traces three tightly coupled ‘critical cases’ that sit at successive system layers. At the *input level*, a forensic audit of privacy-branded *synthetic-data* workflows shows how nominal anonymity can still magnify mis-representation. At the *model level*, a scenario-based study of *machine unlearning* reveals that participants’ wish to be erased is conditioned by collective-risk calculations, not by autonomy alone. At the *evaluation level*, a meta-analysis of *contextual benchmarks* demonstrates that large language models which pass instruction-following tests routinely fail when placed in norm-conflict or long-context settings.

Across these settings the findings converge on four recurring failure modes. First, individual row-level privacy gains often intensify group harms, exposing relational blind-spots. Second, generative and unlearning tools rupture data lineage, undermining regulatory oversight that depends on provenance. Third, performance that looks robust in context-free scorecards collapses under realistic framing, revealing deep contextual brittleness. Finally, the operational cost of verifying compliance with individual-control mechanisms creates enforcement frictions that outstrip regulatory capacity.

Methodologically, the work illustrates how combining dataset forensics, interpretive user research, and quantitative stress-testing produces a *multi-level socio-technical diagnosis* unavailable to any single approach. In doing so it heeds recent calls for a richer ‘evaluation science’ in AI ethics and charts a research agenda toward lineage-aware tooling, context-sensitive benchmarks, and governance structures capable of addressing the relational nature of machine-learning harm.

nunca achei que ia devagar, mas, devagarinho, cheguei
te amo sempre

Contents

Contents	ii
List of Figures	iv
List of Tables	v
1 Introduction	1
1.1 Why revisit privacy’s workhorse paradigm?	1
1.2 Three empirical probes	3
1.3 Contributions in detail	4
1.4 Road map of the dissertation	5
2 Related Work	7
3 Methods	17
4 Case I: Risks of Synthetic Data	22
4.1 Context	22
4.2 Introduction	23
4.3 Related Work	25
4.4 Methods	32
4.5 Risk 1: Diversity-Washing	33
4.6 Risk 2: Circumvented Consent	36
4.7 Discussion	40
4.8 Conclusion	42
5 Case II: Perceptions of Machine Unlearning	44
5.1 Context	44
5.2 Introduction	46
5.3 Related Work	48
5.4 Method	51
5.5 Results	56
5.6 Discussion	61

5.7	Conclusion	66
6	Case III: Contextual Evaluations	67
6.1	Context	67
6.2	Introduction	69
6.3	Related Work	70
6.4	Methodology	72
6.5	Benchmarking Instructions in Context	74
6.6	Benchmarking Contextual Norms	78
6.7	Benchmarking Rules and Logical Constraints	82
6.8	Benchmarking Context, Framing, and Alignment	84
6.9	Synthesis of Methodological Gaps	88
6.10	Opportunities for Future Research	92
6.11	Conclusion	93
7	Discussion	95
7.1	Restating the Core Tension	95
7.2	Cross-Chapter Synthesis of Failure Modes	97
7.3	Aligning Technical Levers & Accountability Pathways	101
7.4	Future Directions: Evaluation, Design, and Policy	104
8	Conclusion	109
	Bibliography	111

List of Figures

5.1	Decision-Making Reference Figure, Participant U	54
5.2	Base Questions and Resulting scenario modifications	55
6.1	Comparative Overview of LLM Adherence Benchmarks	89

List of Tables

5.1	Participant Demographic Data	52
-----	--	----

Acknowledgments

The most important part of this dissertation comes first: none of it could have happened without a whole cavalcade of folks to whom I’m immeasurably grateful. First, I’m immensely grateful to my committee. Thank you, Deirdre, for always providing scaffolding and grace as I bounced between (too) many projects. Thank you, Jenna, for instilling method as a non-negotiable aspect of how I approach research. Thank you, Niloufar, for the early conversations that shaped my comfort with the grey areas of method and practice between disciplines. Thank you all for your support and wisdom during condensed timelines and dispersed time-zones.

The dissertation itself rests on the contributions of my co-authors and collaborators, both those directly involved in the projects which became chapters, and those working on related ones. Among many others, I’m thankful for Jevan Hutson, Niloofar Miresghallah, Andrew Smart, Colin Doyle, Pelle Tracey, Seth Lazar, Helen Nissenbaum, Jessica Newman, Brandie Nonnecke, Shayne Longpre, Anka Reuel, Tobin South, David Piorkowski, Michael Hind, John Roberts, Lilly Irani, and in particular, Justin Norman — our collaborations really feel like a personification of what the I School can be on its good days.

I also thank UC Berkeley’s Center for Long Term Cybersecurity AI Policy Hub for its support, and I acknowledge that this research was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1752814. As is true for so many others, the NSF grant made this possible; the PhD simply wouldn’t have happened without it. Thank you as well to my many study participants, whom I hope I didn’t confuse too much.

So many members of my I School and Berkeley communities also need shout-outs. I’m so grateful for my cohort — especially Tonya, Seyi, and Liza — without whom I can’t imagine having gotten through the first couple years of COVID PhD-ing. Thank you, Emma and Zoe, for being there through so many different confusions and anxieties, and thank you, Inessa and Catherine, for ensuring those confusions and anxieties were always deftly handled. Thank you, Jake, Jevan, and Coye, for so many of the building blocks your classes provided. Thank you, Chase, Simone, Friday, and Tuesday, for the constant support and laughter, and for being the best source of animal companionship on this coast.

Thank you, Lauren, for the copy-editing these last few weeks, the constant support, and for the community you introduced me to. This couldn’t have looked the same without you — I’m lucky to have had a partner who understood the rigmarole of grad school and who, even more specially, cares about the same things. Thank you to my housemates (especially Peter and Sophia for the coffee I’m currently drinking), I can’t imagine having done this without such a deep sense of having found a home, and I’m sorry that it means that AI conversations get brought into the house. Thank you, Alyssa and the Tates, for always being there when I needed you, and thank you, Katie and Brian, for the comfort and escapes. Finally, thank you to my family for not countenancing a world without curiosity. I am so grateful for you all!

Chapter 1

Introduction

1.1 Why revisit privacy’s workhorse paradigm?

Machine-learning (ML) systems shape people’s lives without ever needing a thick biographical dossier on any one of us. Credit scoring, ad targeting, predictive policing, medical triage, and large-language-model chatbots function by detecting *statistical regularities* in vast data sets: correlations between neighbourhood and repayment risk, co-occurrence of words and political ideology, or pixel clusters and racialised identity signals. Pre-computer practices such as red-lining worked in a similar group-based fashion, but today’s ML automates and massively scales that logic. As Viljoen puts it, the governing logic of contemporary data science is *relational* [263]: any person’s fate is inferred from her similarity to reference groups, not from the singular facts of her own life. This insight extends Bowker and Star’s classic analysis of categories as tools of social ordering [38]: when algorithmic systems generalise, they do so by positioning each data subject inside ever finer lattices of collective resemblance.

Yet most legal regimes for data protection are built on a paradigm akin to the funhouse mirror image of that relational reality. U.S. sectoral laws, from HIPAA to FERPA to the Fair Credit Reporting Act, all ask whether an *individual* has been properly informed and whether that individual has consented to a clearly stated use [246]. The European General Data Protection Regulation (GDPR), while broader in scope, implements its “fundamental rights and freedoms” mainly through individual rights of access, rectification, objection, restriction, portability, and erasure [221]. Dan Solove famously calls this architecture *privacy self-management* [247]: by giving each data subject granular notice and choice, the law supposedly enables market and social forces to equilibrate privacy with innovation. Because the harms addressed are framed as losses of *personal* autonomy, the solutions aim to restore autonomy to *that same person*.

A vast technical praxis aligns with the legal one. *Differential privacy*¹ promises mathematical insulation for each row in a data set [81]. *Synthetic-data*² generators advertise ‘privacy-preserving’ swaps because no record maps one-to-one onto a real person [152]. *Machine unlearning*³ offers post-hoc erasure of a subject’s imprint [37]. *Watermarking*⁴ seeks to tag the lineage of data that *remains* in the model [168], while *model deletion orders*⁵ require firms to discard entire models when the underlying records were obtained deceptively [141]. All presume that if the individual regains the power to say “no”, the bulk of the ethical debt is paid.

Mulligan and Doty’s “Design Wars” case study of Apple v. FBI argues that engineering for a *single* public value (in that dispute, device security) can unintentionally ratchet tensions with neighbouring values such as individual privacy, collective safety, and democratic accountability [188, 187]. The very act of hard-coding one solution space closes off others and redistributes power over what counts as a ‘win’. I extend that insight to ML governance: technical fixes that restore autonomy on a *individual, row-by-row* basis may, by design, suppress or externalise harms manifest at group or systemic levels.

This dissertation treats individualized algorithmic remedies — synthetic data, machine unlearning, watermarking, and model deletion — as *stress tests* for the notice–consent–erasure paradigm itself. Each tool is explicitly premised on the idea that autonomy can be restored one data subject at a time. By tracing how these remedies perform in real technical, organizational, and social contexts, I ask two guiding questions:

RQ1 — Internal efficacy. When notice–consent–erasure mechanisms are implemented end-to-end, how much *operational control* do data subjects actually gain, and where do those mechanisms falter or backfire?

RQ2 — Residual and systemic harms. After those limits are exposed, which relational, systemic, or representational harms (i) still persist, (ii) are actually *worsened*, or (iii) are displaced into new design arenas, and how are measurement practice and governance required to surface and mitigate them?

The next three chapters constitute empirical probes aimed squarely at these questions.

¹Guarantees that adding or removing a single record changes an analysis by at most e^ϵ ; see Section 2.2.

²Artificially generated records that imitate the joint distribution of real data; origins and desiderata reviewed in Section 2.2.

³Algorithms that retroactively erase a data point’s influence on a trained model; background in Section 2.2.

⁴Techniques that embed cryptographically detectable signals in model parameters or outputs so that downstream uses can be traced back to a specific training corpus; details in Section 2.2.

⁵Regulatory mandates, most notably from the U.S. Federal Trade Commission, that require companies to destroy both unlawfully collected data and any models or derivatives trained on that data; see Section 2.2.

1.2 Three empirical probes

1. Synthetic data as lineage-scrubbing and diversity-washing (Chapter 4)

Synthetic data is advertised as a two-for-one fix: mask direct identifiers to placate privacy law, and generate synthetic examples that mimic individuals from under-represented sub-populations to placate fairness audits. Using face-recognition pipelines as a concrete setting, Chapter 4 demonstrates:

- **Lineage opacity.** Once real and synthetic images are blended, it becomes almost impossible — even for the dataset creator — to trace which real faces seeded which augmented clusters⁶. FTC investigators would have no hook to prove deceptive data collection; a prerequisite for model deletion orders [85].
- **Designer power.** Procedural generation tools (e.g. Synthesis AI, FaceSynthetics) permit the designer, not the depicted communities, to choose which phenotypes get rendered and in what frequency. The pretence of ‘statistical parity’ masks profound gaps in cultural or contextual fidelity [148].
- **Diversity-washing.** Benchmark accuracy rises because *tail identities* — rare classes such as darker-skinned women over sixty — are deliberately oversampled, yet the synthetic faces lack real-world texture and lighting and they shift class priors, so the model learns brittle shortcuts and looser thresholds; when those artefacts disappear in deployment, false-positive rates for the same marginalised groups *also* climb, exemplifying Ruha Benjamin’s “New Jim Code” dynamic: technical fixes that promise fairness while reinserting hierarchy [28, 123].

2. Machine unlearning and the social calculus of erasure (Chapter 5)

If machine unlearning were simply a drop-in compliance knob that makes full erasure technically and economically routine, one might expect data subjects to invoke it whenever their privacy preferences tighten. A scenario-based elicitation with 21 employees of a global technology firm paints a more layered picture:

- Participants weigh **domain stakes**. For instance, they treat housing or health risk scores as arenas where even tiny data leaks warrant drastic erasure, yet shrug off similar leaks in a food-delivery context.

⁶See Chapter 7 for a discussion of paths forward to mitigate this.

- They practise **gap-motivated reasoning**. Faced with opacity about model logic, some default to optimism (“outcomes matter, not internals”); others default to suspicion. Either position might flip after a breach.
- They consider **downstream fairness**. Several fear that deleting or unlearning only their own record could mathematically worsen bias against peers — an altruistic concern beyond the individual-control script.

Thus, even a perfect unlearning primitive cannot escape the structural bind it creates: the individual-control frame forces people to weigh personal privacy against the potential harms their absence may impose on others, leaving distrust, power asymmetry, and group-level effects unresolved.

3.Contextual evaluation and the individual-control lens (Chapter 6)

Our survey of thirty-three adherence benchmarks shows that most inherit, in miniature, the very logic of *individual control* critiqued throughout this dissertation. Each prompt is handled as its own micro-contract: the model is judged on whether it fulfils that *single* instruction, safety cue, or rule, with no reference to what came before, who else is speaking, or which social norms the setting activates. Less than a third of the benchmarks provide any narrative context; only three deliberately create a conflict between the prompt and a surrounding norm; and none measure how performance shifts when the same request is embedded in a different cultural, role-based, or long-context frame.

The consequence is a familiar blindness. A model that aces IFEval’s keyword checks stumbles on 40 % of NormAd vignettes once the etiquette is Ghanaian rather than U.S., and the same model that passes single-turn safety filters releases prohibited medical advice after the cue is buried mid-context in LongBench. These breakdowns mirror the fractures seen earlier in the pipeline: synthetic data protects rows while distorting group representation (Chapter 4); machine unlearning respects an individual request while leaving cohort fairness at risk (Chapter 5). In evaluation, the analogue is prompt-by-prompt scoring that treats appropriateness as detachable from *who* speaks, *to whom*, and *for what purpose*. Chapter 6 therefore extends the dissertation’s main claim to the utilisation layer: as long as measurement stays bound to an atomised, one-prompt view of control, relational and contextual harms will remain invisible — and ungoverned — in deployed LLMs.

Evaluation pipelines, like legal ones, remain calibrated to the individual-control lens.

1.3 Contributions in detail

The dissertation makes five concrete contributions, spanning empirical evidence, qualitative insight, critical synthesis, and policy design:

1. **Empirical evidence of lineage, power, and fairness risks in synthetic-data practice.** Through reproduction studies, Chapter 4 documents how synthetic pipelines can simultaneously satisfy privacy formalism and deepen representational harm.
2. **A qualitative study of when and why data subjects invoke unlearning.** Chapter 5 presents scenario-based interviews with 21 technology professionals uncover four recurring considerations — personal agency, domain stakes, downstream fairness, and institutional trust — that govern when people choose opt-out, deletion, or machine unlearning, showing that concerns for collective impact often override a purely individual-privacy calculus.
3. **A critical synthesis of context-aware evaluation research.** Chapter 6 maps where current benchmarks fail to capture instruction-norm interference, normality manipulation, and long-context drift, and it outlines desiderata for next-gen ‘interference-aware’ tests.
4. **An integrated argument for supplemental, collective safeguards.** Across the three studies, the dissertation demonstrates that the notice-consent-erasure paradigm cannot resolve many relational harms, and proposes high-level policy and design shifts — namely lineage-preserving synthetic data standards, combined erasure & fairness audits, and participatory dataset governance — to complement individual rights.
5. **Policy implications.** Concrete recommendations for regulators include:
 - (a) require lineage-preserving synthetic-data pipelines,
 - (b) pair unlearning mandates with collective fairness audits, and
 - (c) extend model deletion remedies beyond deceptive collection to encompass certain classes of relational harm.

1.4 Road map of the dissertation

Chapter 2 reviews the intellectual history of privacy self-management, relational theories of data, and sociotechnical harm taxonomies.

Chapter 3 sets out a comparative multi-case, critical-sociotechnical methodology — combining dataset forensics, scenario-based elicitation, and benchmark-landscape analysis — to trace how individual-control logics break down across data, model, and evaluation stages.

Chapter 4 evaluates synthetic-data pipelines in facial recognition and unmasks lineage, bias, and power implications.

Chapter 5 reports the scenario-based study of machine unlearning and RTBF decision-making.

Chapter 6 synthesises benchmark literature and proposes new interference-aware evaluation directions.

Chapter 7 discusses RQ1 and RQ2 and, drawing on Mulligan and Bamberger [187], shows how single-value ‘fixes’ can collide with security, equity, or democratic oversight when harms scale beyond the individual.

Recognizing the limits of individual control is not to discard autonomy; rather, it is to situate autonomy inside the relational webs that contemporary machine learning relentlessly exploits. Only by mapping where the self-management notice-consent-erasure paradigm fractures can we craft the collective, context-sensitive protections that twenty-first-century data subjects — and the societies they compose — actually require.

Chapter 2

Related Work

2.1 Intellectual Lineage of the Individual-Control Paradigm

The notion that privacy rests on *individual control* over personal information originates in U.S. legal thought at the end of the nineteenth century. Warren and Brandeis framed the ‘right to privacy’ as the right to decide “when, how, and to what extent” facts about oneself may be made public [270]. Mid-century tort doctrine kept that individualist lens, enumerating disclosure, intrusion, appropriation, and false-light as personal causes of action [211]. Alan Westin’s landmark 1967 monograph translated the principle into the language of information systems, defining privacy as the individual’s authority to “control, edit, manage, and delete” data about herself [274]. Together these works established autonomy and choice as the conceptual core of privacy.

Computers in government and business spurred a policy turn. The 1973 U.S. Department of HEW report set out the five *Fair Information Practices (FIPs)*: notice, consent, access, integrity, and security [262]. A companion inquiry in the United Kingdom, the 1972 Younger Committee report, recommended essentially the same principles [65]. On the Continent, the German state of Hessen went further, enacting the world’s first data-protection statute in 1970 (revised 1978) and installing an independent privacy commissioner — an institutional design traced in Mulligan’s Germany chapter of *Privacy on the Ground* [134, 19]. The 1980 OECD Guidelines knitted this trans-Atlantic consensus into an international template; harmonised FIP-style obligations as the price of unrestricted data flows [199]. Subsequent U.S. sectoral statutes (HIPAA, COPPA, GLBA) and the European Union’s GDPR inherit the same individual-rights architecture, but with an important divergence: the GDPR treats consent as only one lawful basis alongside contractual necessity, legitimate interest, purpose limitation, and data minimisation, preventing it from becoming the ‘get-out-of-jail-free’ card it often is in U.S. practice.

By the mid-1990s, critiques of the liberal autonomous self at the centre of privacy protection were gaining ground. Regan argued in 1995 that privacy also serves collective ends —

democratic participation and group identity — not protected by atomised transactions [223]. Solove diagnosed the ‘consent dilemma’: practical limits on users’ time, comprehension, and power render meaningful choice illusory in large-scale data markets [247]. Behavioural studies confirmed a persistent ‘privacy paradox’, whereby users disclose despite stated concern, trapped by structural asymmetry rather than personal inconsistency [24]. At the same time, revelations of state-sponsored mass surveillance [122] and critical scholarship on big-data epistemology [40] showed that informational asymmetry was not a bug but a feature of modern data governance.

Scholars then re-conceptualised privacy as relational and contextual. Nissenbaum’s theory of *contextual integrity* assesses the propriety of data flows against shared social norms rather than individual preference [196]. Doty, Koopman, and Mulligan demonstrated that privacy is an “essentially contested concept” in technical standard-setting, revealing deep normative disagreements that surface even in engineering fora [189]. Bamberger and Mulligan later showed empirically that companies operationalise markedly different privacy logics in practice [19]. Viljoen made the case that data are constitutively *relational*: their very value arises from patterns across people, so harms likewise manifest at supra-individual levels [263]. Complementary critiques expose the political-economy stakes: Cohen’s “Between Truth and Power” charts how informational capitalism reconfigures both market power and legal accountability [63], Zuboff’s ‘surveillance capitalism’ frames consent as a coerced economic transaction [294], and Floridi’s ‘group privacy’ surfaces harms borne by collectives rather than persons [104].

Taken together, this intellectual lineage explains both the persistence and the fragility of the notice-and-consent paradigm. It codified autonomy at a moment when large-scale data processing was still exceptional and partly transparent: information flows were centralised in governments and a handful of firms — the ‘big three’ credit bureaus, major insurers, national census agencies — and these actors typically had to ask individuals to supply data, creating at least the appearance of participation. Because the primary policy aim was to prevent illegitimate *identification*, data-protection law never set out to police *representational* or statistical harms; indeed, aggregate knowledge was viewed as both rational and beneficial [108]. Only once pervasive, low-friction surveillance let organisations single people out without ever re-identifying them did those harms become visible — prompting targeted interventions such as the U.S. Genetic Information Non-discrimination Act (GINA, 2008), the first statute to treat predictive statistical profiles themselves as a source of harm [112]. That historical tension motivates this dissertation’s empirical probes into synthetic data, machine unlearning, and context-blind evaluation, each of which shows how technical fixes anchored in individual control falter when confronted with the relational realities of contemporary machine learning.

2.2 Technical Artefacts that Operationalise Individual Control

The legal logic of notice-and-consent inspired a complementary suite of technical mechanisms that promise to engineer individual control into machine-learning (ML) pipelines. These artefacts cluster into four overlapping families:

- **Privacy-preserving statistics**, such as differential privacy, which bound the risk that any *one* record can be *singled out* from a released statistic [81, 61];
- **Data substitution**, typified by synthetic data, which replaces raw rows with statistically similar proxies;
- **Post-hoc revocation**, e.g. machine unlearning, which seeks to erase an individual’s influence by updating the model after training; and
- **Traceability and deletion guarantees**, including watermarking regimes and model deletion orders, which aim to prove (or legally compel) the deletion of models built on unauthorised data.

All four families pursue privacy by shrinking or nullifying a single record’s influence, thereby reinforcing the premise that privacy is a per-subject entitlement rather than a relational property.

Differential privacy

Differential privacy (DP) guarantees that the *observable distribution* of any analysis changes only minimally — by at most a multiplicative factor e^ϵ — when a single record is added or removed [81]. The promise is not absolute ‘safety’, but a mathematically bounded privacy loss that lets an individual participate in a database without facing disproportionate new risk.

Differentially-Private Stochastic Gradient Descent (DP-SGD) extended that guarantee to deep learning by injecting calibrated noise at each gradient step, enabling production use at Google and Apple [1]. Rényi differential privacy tightened the calculus for composition and sharper bounds [185], while Abowd and Schmutte analysed the privacy-accuracy trade-off from a social-welfare perspective, quantifying the costs of large-scale deployments [4]. To reconcile utility with protection, Kohli et al. show how domain-specific statistical targets can guide privacy-budget allocation, preserving task accuracy while up-holding stringent ϵ values [160].

Even so, memorisation attacks on ostensibly private models persist [51]. Budgets low enough to guard individuals can cripple downstream accuracy, tempting deployers to choose larger ϵ settings that erode the original autonomy guarantee. The gap between formal protection and practical deployment thus remains a live zone of contestation — one this

dissertation returns to when assessing lineage opacity in synthetic data and revocation limits in machine unlearning.

Synthetic data

Where differential privacy tries to hide a person ‘in the crowd’ by adding noise, synthetic data takes a more radical step: it fabricates new records so the real ones never leave the vault. A generative model — first generative adversarial networks (GANs) [117], now also diffusion models and specialised tabular generators — learns the statistical relationships in the original dataset and then produces artificial samples that *look* plausible but (ideally) do not correspond to any actual individual.

In principle, this surrogate data can be used for the same downstream tasks — model training, dashboarding, exploratory analysis — while sidestepping the legal and ethical burden of sharing raw personal information. Jordon et al. frame the design triad as *utility* (how well the synthetic data supports the task), *fidelity* (how faithfully it matches the real distribution), and *privacy* (how hard it is to trace a synthetic record back to a real person) [152].

Reality is messier. Many deployments use synthetic data to *augment* rather than fully *replace* original records — for example, inflating rare classes in face-recognition benchmarks. Utility often wins out over lineage: Atzori et al. show that fairness claims based on synthetic ‘diversity boosts’ collapse under rigorous distributional audit [15]. Worse, if the generator memorises its training set, fragments of the original images or rows can leak into the synthetic output, recreating the very individual harm the technique was meant to avoid [51].

The chapter that follows (Chapter 4) probes these trade-offs in depth, asking whether synthetic pipelines can truly deliver both statistical usefulness and verifiable privacy — or whether, in practice, they trade one kind of risk for another.

Machine unlearning

Consent in modern data law is not a one-time event. Under the GDPR’s *right to erasure* (Art. 17) and a growing patchwork of U.S. state statutes, individuals may later demand that their personal data be deleted *and* that systems cease relying on it. The texts do not prescribe a specific technical pathway for meeting that demand, but they raise an obvious practical question: once information has been baked into a model’s parameters, what counts as ‘forgotten’? **Machine unlearning** is the research community’s leading attempt to answer that question.

The basic idea is simple to explain — even if the algorithms are not. Instead of retraining a full model from scratch each time a record must be erased, an unlearning procedure updates the existing model (or a small slice of it) so the erased record no longer influences the outputs. Bourtole et al.’s SISA framework showed one practical path: shard the training set, retrain only the affected shard, and then re-aggregate the shards [37]. Since then, researchers have

proposed exact, approximate, and formally certified techniques; Liu et al. map this fast-growing landscape [172] ¹.

Making the remedy work at scale is tricky. Faster unlearning often means noisier gradients and poorer calibration, especially in deep-learning pipelines; Neel et al. document the resulting utility–time trade-offs [192]. In other words, the tool can restore an individual’s autonomy but at a collective cost — slower inference, lower accuracy, or bias drift. Chapter 5 investigates that trade-off empirically and through scenario-based user studies, asking when people actually *want* the remedy and what side-effects they are willing to tolerate.

Traceability and mandated deletion

Both preventive tools (differential privacy, synthetic data) and revocation tools (machine unlearning) depend on some form of *provenance*: to remove a record — or prove it was never there — you must be able to trace which inputs shaped which outputs. Traceability technologies try to make that linkage explicit.

Watermarking is the most visible technical line. During training the model’s weights are stamped with a cryptographic signature tied to a specific dataset; if an unauthorised copy surfaces, the watermark can reveal whose data — or whose licensing terms — were violated [158]. Audit logs and secure enclaves push in the same direction, maintaining tamper-evident records of which files, features, or gradients entered each training run.

Regulators are beginning to weaponise provenance. Since 2021 the U.S. Federal Trade Commission has issued a series of ‘model deletion’ orders (*Everalbum*, *Rite Aid*, *Ring*, *Clearview*) that compel firms not only to delete unlawfully obtained data but also to destroy the *models* and *derivative products* built with it [141]. The burden of proof falls squarely on the company: if it cannot document which data touched which model components, the safest legal course is to delete everything.

That inversion of risk matters for the artefacts studied in this dissertation. Synthetic-data pipelines sever clear lineage, making selective deletion almost impossible. Unlearning relies on logs that many production teams discard to save storage. Watermarking itself breaks under model compression or noisy fine-tuning. In short, the stronger the legal remedy, the more brittle the supporting provenance infrastructure — an asymmetry Chapter 4 and Chapter 5 expose in detail.

Synthesis

Each technical artefact reviewed above inherits a *row-centric threat model*: a single row in a table corresponds to a single person, so *harm* is defined as the ability to draw an unwanted inference about **that** person, and *success* is defined as bounding or erasing **that row’s** influence. By design, the broader relational harms surveyed in Section 2.3 — biased

¹As does forthcoming non-dissertation work of mine with Jevan Hutson and Jay Conrad!

representation, allocative skew, and systemic feedback loops — fall outside the protection scope.

The empirical chapters that follow put those boundaries under stress. Chapter 4 shows how synthetic-data pipelines, while obscuring lineage for any one individual, enable diversity-washing that distorts group representation. Chapter 5 demonstrates that machine unlearning delivers per-record revocation yet destabilises collective utility and fairness. Chapter 6 reveals that benchmark culture, focused on prompt-level correctness, rarely detects multi-subject interference or population-level degradation.

Taken together, the findings point to a structural limit: technical mechanisms that faithfully implement the individual-control paradigm are intrinsically ill-suited to address *relational* harms — statistical discrimination, misrepresentation of entire communities, and feedback loops that entrench inequality. Closing that gap will require measurement practices and legal standards that look beyond the row, toward groups, contexts, and infrastructures.

2.3 From Individual Wrongs to Relational Harms

Over the past decade, empirical research and critical theory have shown that privacy violations are only one slice of the harms enabled by large-scale personal-data processing. Long before today’s platforms, Gandy’s *Panoptic Sort* warned that statistical profiling would naturalise ‘rational discrimination’ across credit, insurance, and employment sectors [108]. Recent scholars of race, gender, and political economy trace that logic through modern AI systems: Benjamin’s *New Jim Code* reveals how face-analysis and predictive-policing tools capture Black bodies with disproportionate intensity — over-represented in surveillance footage yet *under*-represented in training data — then deploy higher error rates to justify punitive oversight [27]. Noble demonstrates that misogynoir can arise from click-through statistics alone, without any explicit demographic fields [197]. Eubanks follows welfare-screening algorithms that transform poverty itself into a risk signal, converting need into bureaucratic punishment [90]. In all these cases the harm is not mere data leakage but the *discriminatory classification* of social groups; a structural effect that neither notice nor after-the-fact revocation can remedy.

Parallel theorists argue that contemporary data flows amount to resource extraction aimed at collectives. Couldry and Mejias call this ‘data colonialism’, in which everyday human experience is mined through coerced digital labour [67]. Surveillance-studies scholars reach a similar verdict: Lyon’s ‘dataveillance’ and Haggerty & Ericson’s ‘surveillant assemblage’ frame population-level sorting as the dominant logic of twenty-first-century control [177, 127]. Floridi’s notion of ‘group privacy’ adds a normative kicker: algorithmic inference creates rights-bearing entities — ethnic communities, gender identities, gig-worker cohorts — that the standard consent paradigm simply cannot protect [104]. These perspectives converge with Viljoen’s relational theory of data governance, swapping the row-wise ontology of databases for a column-wise vision of interdependence. Privacy thus becomes a question of how populations are sorted and steered, not merely of who clicks ‘agree’.

Taxonomies of sociotechnical harm make that abstraction empirically tractable. Shelby et al. distinguish representational, allocative, quality-of-service, interpersonal, and systemic harms, placing conventional privacy infractions in only one of five categories [237]. Concrete studies fill in the grid: Benthall and Keyes show how racial categories hard-coded in vision benchmarks propagate exclusion [29]; Andrus et al. demonstrate that ‘demographic-reliant fairness’ reproduces historic subordination when labels are missing or contested [10]; Birhane and Prabhu trace non-consensual images from the web into large-scale training sets, detailing a cascade of *dis-representation* — the systematic erasure or denigration of certain groups rather than simple mis-representation [31]. Even synthetic data, long touted as a privacy panacea, amplifies phenotype bias and cultural erasure: Mei et al. show that GAN (and diffusion) generated faces skew lighter and more Euro-centric than their source distributions [125]. These findings reveal relational harm in action: individuals cannot ‘opt out’ of discriminatory classification that is baked into the upstream data pipeline.

Recognising such harms has led scholars to propose governance mechanisms no longer centred on atomised consent. Gürses and van Hoboken call for *collective data governance* that treats datasets as public infrastructures, subject to fiduciary duty rather than contract [126]. Ajunwa argues for due process guarantees and disparate impact testing in algorithmic hiring, borrowing tools from civil rights law [8]. Citron and Solove extend privacy torts to cover emotional and social injuries, acknowledging non-economic dimensions of harm [60]. Selbst et al. caution that purely technical ‘fairness fixes’ often abstract away contextual nuance, calling for interdisciplinary design that embeds social meaning into ML pipelines [233]. Collectively, this body of work insists that protecting autonomy requires systemic remedies — audits, participatory benchmarking, communal veto — not merely better notice or reversible training.

For the dissertation that follows, these critiques establish the evaluative baseline: a technical solution that satisfies individual control yet leaves relational harms intact is inadequate. By analysing synthetic data and machine unlearning through the lens of lineage opacity, diversity-washing, and collective trust, the subsequent chapters test whether individual-centric tools can ever grow to meet relational obligations, or whether new paradigms are required.

2.4 Measurement and Enforcement: The Per-User Gaze

Quantitative evaluation in machine learning has long mirrored the notice-and-consent paradigm by treating the *prompt* or *record* as the unit of analysis. Early large-language-model (LLM) benchmarks such as IFEval test whether a model obeys keyword, length, or format constraints that can be verified automatically for each response [292]. Success is binary: either this answer respects *my* instruction or it does not. Subsequent work expanded complexity — ComplexBench composes multiple constraints [273] and FollowBench adds

them level by level to observe degradation [150] — but the granularity remained individual. These designs inherit two assumptions embedded in the Fair Information Practices (FIPs): that conformance is measurable at the point of interaction, and that each interaction is independent of all others.

A safety turn in LLM research introduced new metrics but retained the same granularity. RealToxicityPrompts probes single-turn toxicity [111]; BIG-Bench grades hundreds of independent tasks with a per-prompt rubric [250]. More recent efforts recognise context yet still evaluate one user at a time. CASE-Bench uses Contextual Integrity to frame *safe* vs. *unsafe* ask-context pairs, but ground-truth labels are derived from crowd judgments on each isolated query [252]. FaithEval, designed for retrieval-augmented generation, measures whether a single answer is entailed by its retrieved snippets — not whether the retrieval itself reflects group bias [184]. ContextualJudgeBench goes a step further by testing whether GPT-4 can grade answers that depend on external context, yet again the metric is a point-wise agreement score [284]. What remains unmeasured is interference: How does satisfying one user’s instruction alter harms experienced by another? How does refusal in a safe context propagate over a multi-user conversation?

Legal enforcement largely tracks the same per-subject logic. The EU GDPR’s erasure right (Art. 17) empowers individuals to compel deletion of their records. U.S. regulators rely on the same lineage: Illinois’ Biometric Information Privacy Act (BIPA) creates a private right of action for un-consented biometric capture, and the California Consumer Privacy Act (CCPA) offers opt-out and deletion on a per-consumer basis [93, 143, 47]. The U.S. Federal Trade Commission’s model deletion orders in *Everalbum*, *Rite Aid*, and *Ring* required deletion of models trained on images or videos collected without valid consent; in *Rite Aid*, the complaint additionally cited systemic bias against shoppers in “Black or African-American neighbourhoods”, yet the remedy still focused on purging the identifiable victims rather than repairing the model’s disparate impact [97, 98, 95]. Even the FTC’s path-breaking Clearview settlement, which forces the deletion of all face vectors scraped from certain platforms, grounds its authority in deceptive acquisition, not in representational or systemic harm [96]. The forthcoming EU AI Act does ban real-time remote biometric identification and certain forms of indiscriminate surveillance, but its high-risk tier remains keyed to systems that directly affect ‘natural persons’, leaving wide-scale population modelling subject mainly to transparency and audit duties [92].

Practical frictions compound the conceptual blind spot. Synthetic-data pipelines sever auditable lineage: after an image is ‘re-imagined’, no pixel maps cleanly back to a consent form. Certified unlearning schemes must either retain logs sufficient to prove compliance — re-introducing privacy risk — or settle for unverifiable best-effort guarantees [172]. Watermarking promises provenance but breaks under model compression or noisy fine-tuning [158]. In FTC practice, investigators cite not only uneven documentation but also resource asymmetries, opaque third-party vendors, and live model updates that outpace legal-hold orders, forcing reliance on company affidavits over technical proof [141]. All of these hurdles flow from a per-record theory of harm that demands a per-record accounting infrastructure — an infrastructure that collapses under contemporary ML supply chains.

For this dissertation, such limitations clarify why technical ‘fixes’ succeed in benchmarks or courtrooms yet fail in social reality. Chapter 4 shows that lineage loss in synthetic data undermines both utility and regulatory traceability; Chapter 5 illustrates how unlearning satisfies individual erasure while destabilising group-level fairness; Chapter 6 demonstrates that point-wise metrics miss the context collisions driving representational and systemic harms. Echoing Mulligan’s “Design Wars” critique that optimising for a single value can imperil others [187], the larger agenda here is to re-orient measurement and enforcement toward collective outcomes — an ambition unreachable so long as benchmarks and laws stare, record by record, through a notice-and-consent lens.

2.5 Synthesis and Research Agenda

Taken together, the four strands of scholarship reviewed above map a clear but uneven terrain. Legal theorists and policy architects have equipped us with a coherent doctrine — privacy as self-management — that still structures statutory rights, corporate playbooks, and benchmark design (2.1). Computer-science research has refined that doctrine into deployable artefacts that can mathematically bound or retrospectively revoke an individual record’s influence (2.2). Critical STS, race, and data-governance scholars, however, reveal a second topography in which harm propagates through representational distortion, allocative skew, and infrastructural asymmetry — phenomena that are not merely left out of the individual-control frame but in some cases actively intensified by it (2.3). Evaluation culture and enforcement mechanisms then reinforce the first topography by measuring success one prompt, one row, one litigant at a time, leaving the relational landscape largely uncharted (2.4).

This bifurcation between the individual and the relational produces three interlocking gaps that motivate the empirical chapters which follow. *First*, there is a descriptive gap: we lack fine-grained evidence of how tools engineered for per-record privacy reconfigure group-level outcomes in real ML workflows. Synthetic data, machine unlearning, and instruction-following benchmarks are lauded in isolation but rarely examined for collateral effects on bias propagation, collective trust, or auditability. *Second*, an explanatory gap persists: existing taxonomies of sociotechnical harm do not yet specify the causal pathways by which ostensibly privacy-enhancing interventions yield diversity-washing, calibration drift, or lineage opacity. *Third*, there is a governance gap: current regulatory triggers — deception, consent defect, biometric capture — address only a subset of harms surfaced by relational critiques, leaving regulators with blunt or mis-targeted tools when collective injuries emerge.

The dissertation addresses these gaps through a sequential research design that treats each technical artefact both as an *object of study* and as a *lens on the paradigm that produced it*. Chapter 4 scrutinises synthetic data pipelines in the context of computer-vision benchmarking, combining statistical audits with qualitative analysis of dataset-builder rationales. Chapter 5 turns to machine unlearning, pairing scenario-based user studies with performance experiments to reveal how revocation rights collide with model calibration and

social trust. Chapter 6 interrogates evaluation culture itself: by surveying benchmarks that encode context, instruction adherence, or safety norms, the chapter shows how relational harms are elided at the metric definition stage. Across all three sites, the analytic strategy is to follow a single question: *What happens to relational harm when we maximise individual control?*

The intended contribution is twofold. The dissertation demonstrates that solutions praised for their privacy guarantees can generate second-order effects — diversity-washing, lineage invisibility, over-broad refusals — that routinely elude both benchmark detection and statutory remedy. Conceptually, this study reframes privacy engineering as a socio-technical trade-off space rather than a one-directional progress curve, echoing Mulligan’s “Design Wars” insight that optimising for a single value can endanger others [188, 187]. By detailing where and why the individual-control model proves incomplete and inadvertently exacerbates other harms, the dissertation lays groundwork for alternative governance logics such as collective consent, fiduciary obligation, and context-based risk assessment.

In closing, the chapters do not seek to discard individual control — autonomy remains normatively vital — but to situate it within a broader constellation of relational values that need protection from the harms machine-learning systems intensify. Only by acknowledging both vantage points can ML governance move from row-wise guarantees toward genuinely systemic accountability; a trajectory the remainder of this dissertation now sets out to trace in practice.

Chapter 3

Methods

3.1 Analytic Framework and Evidence Strategy

Machine-learning systems are never *just* code; they are crystallisations of labour, politics, and institutional bargains. This dissertation therefore begins from a *critical sociotechnical* stance that treats technical artefacts and social arrangements as co-produced. Feminist STS reminds us that all knowledge is *situated*: every chart, benchmark, or API call is issued from a particular vantage, constrained by policies, institutions, and the designer’s material accountabilities [130]. In parallel, the *critical technical practice* movement in HCI argues that rigorous engineering requires iterative interrogation of the social assumptions baked into design choices — what Agre called “folding critique back into making” [6, 79].

Multi-layer evidence.

Taken together, these lineages license a methodological ethic in which *source code*, *regulatory filings*, and *interview talk* are read as equally consequential traces of the same sociotechnical order.

- **Reading code** exposes the implicit data models, error tolerances, and fail-safe assumptions that shape downstream behaviour;
- **Writing code** as replication reveals what shortcuts are necessary in practice, surfacing ethical trade-offs that rarely appear in publications;
- **Document analysis** of filings, policies, and consent disclosures shows how legal and organisational constraints frame what the system is *asked* to do;
- **Interviews** with engineers, policy staff, and affected users capture the lived frictions that tie technical artefacts to social practice.

Relational lens on harm.

Building on this premise, the project adopts a *relational view of harm*: ethical failures emerge *between* people who share salient features, not solely within isolated data ‘rows’. Handling issues only at the level of the individual is therefore doubly insufficient: (1) it misses inequalities that propagate across group likenesses, and (2) it ignores the fact that many harms materialise only *after* deployment, as feedback loops tighten around real-world usage. Viljoen’s theory of relational data flows [263], Barocas and Selbst’s contextual fairness framework [23], and Dwork and Mulligan’s “It’s not privacy and it’s not fair” critique, echoed by Hoffman [82, 135], all warn that notice-and-consent remedies are blind to these supra-individual dynamics.

Analytic generalisation over statistical universalism.

Because relational harms are emergent and context-bound, the study pursues *analytic generalisation* rather than statistical universalism. Each in-depth case is designed to surface mid-range concepts — such as *provenance opacity* or *contextual brittleness* — that can travel across settings without claiming to exhaust them. The objective is to map the space of failure modes, not to count their global frequency.

In sum, the methods treat ethics, engineering, and governance as a single field of inquiry: ethics supplies the questions, engineering supplies the boundary conditions, and governance supplies the stakes. The chapters that follow operationalise that stance by interleaving code audits, statistical tests, document analysis, and interview-based sense-making.

Empirical Designs in Context

The epistemic commitments above dictate a multi-pronged empirical strategy in which each chapter functions as a deliberately different ‘probe’ into the same individual-versus-relational tension. What follows is a preview of those designs. Full protocols and instruments appear in the respective chapters, but the sketches here are meant to orient the reader by showing how the critical-sociotechnical lens translates into concrete sampling choices, data-collection techniques, and analytic workflows.

Chapter 4 — Practice-Centred Audit of Synthetic Data

Chapter 4 is framed as a *practice-centred audit*. Reverse-engineering a synthetic-data workflow is paired with close reading of FTC consent orders, allowing technical artefacts and legal concepts to be analysed as a single regulatory assemblage. Rather than running a large-N statistical study, we step into the role of an expert practitioner and rebuild a face-recognition evaluation pipeline exactly as it is promoted in vendor white papers and forensic-tech manuals. The workflow uses SYNTHESIS AI to generate a gallery of 50,000

procedurally created faces that the tool advertises as “balanced across age, gender, and skin-tone”. We also inject a probe set built from real CASIA–WebFace images that have been methodically degraded—blurred, JPEG-compressed, and heavily down-sampled—to mimic the low-quality footage on which police and border-control systems increasingly rely. The resulting pipeline looks, on paper, like a textbook privacy-by-design exercise: every probe has been stripped of direct identifiers, and the gallery is numerically ‘diverse’.

Running the evaluation, however, exposes how easily *statistical* diversity can hide a lack of *representational* diversity. ‘Balanced’ output still over-represents light-skinned male phenotypes and offers only slight texture variation for darker-skinned women. The evaluation thus passes row-level privacy checks while silently reproducing the very representational gaps that motivated the switch to synthetic data in the first place. By following best practice and still ending up with a diversity-washed dataset, the case shows how individual-control thinking — ‘as long as each record is de-identified, we are safe’ — obscures the collective and cultural dimensions of harm that synthetic data can amplify.

The second half of the chapter uses two emblematic settlements, *Everalbum* (2021) and *Clearview AI* (2022), to illustrate how U.S. enforcement still depends on a provable chain of custody from model artefacts back to identifiable, non-consented data. Placing those vignettes alongside the provenance-breaking synthetic-data workflow makes the contrast vivid: once real faces are transformed into lineage-free synthetic images, that evidentiary trail disappears, leaving regulators with no foothold for the same remedy. The governance discussion therefore demonstrates, by example, why traceability is a silent prerequisite for current deletion orders and how synthetic pipelines can inadvertently erode that foundation.

By pairing an industrially plausible evaluation with a close reading of enforcement doctrine, Chapter 4 demonstrates that a mechanism engineered to reduce individual privacy risk can at the same time distort group representation and erode the very consent-based levers regulators rely on. The mixed-method design therefore provides the baseline empirical evidence for the dissertation’s broader claim: tools that perfect row-level control often intensify relational harms and blunt accountability.

Chapter 5 — Scenario-Based Elicitation of Machine-Unlearning Decisions

To examine how real data subjects weigh the promise of post-training deletion against utility and fairness concerns, Chapter 5 employs a scenario-based elicitation with twenty-one professionals drawn from a global technology firm. Participants were deliberately heterogeneous in job function and AI literacy — marketing managers, software engineers, designers, researchers, operations, and sales staff — so that discussion could begin at practical trade-offs rather than introductory explanations. Semi-structured ranking tasks elicit how participants reason about unlearning in situated contexts, foregrounding meaning-making over behavioural prediction and enabling grounded theorisation of privacy agency.

Each participant joined a sixty-minute video interview structured around three realistic

services selected from the AI Incident Database: a food-delivery app that targets ads from phone telemetry, an insurance engine that prices policies using credit and public-web data, and a landlord tool that sets rent from smart-home sensors. Within every scenario the interviewer varied five disclosure dimensions — training-data scope, output use, accuracy, decision logic, and third-party sharing — presenting an ‘unknown’, ‘good’, and ‘bad’ version of each. Participants first ranked the fifteen cards from most to least comfortable, then drew three threshold lines indicating the point at which they would (i) opt out entirely, (ii) delete existing data, or (iii) request machine unlearning so the model would behave as if their data had never been included.

All sessions were transcribed verbatim and inductively coded by two researchers using a consensus-coding protocol [182]. The analysis surfaced four cross-cutting considerations — personal agency, domain stakes, downstream fairness, and institutional trust — that organise the findings in Chapter 5 and illustrate why a per-record remedy can still destabilise collective utility and equity.

Chapter 6 — Survey of Context-Aware Benchmarks

To chart how the LLM community evaluates instruction, norm, and rule adherence, I carried out a systematic yet deliberately lightweight literature review spanning January 2020 through March 2025. Searches on ACL Anthology, arXiv, Semantic Scholar, and Google Scholar combined terms for instruction following, social or cultural norms, rules, and benchmarking — queries such as “*instruction following AND benchmark AND LLM*” or “*social norm evaluation*”. A study qualified for inclusion if it (i) reported empirical results for a model exceeding one billion parameters, (ii) introduced or reused a benchmark whose *primary* goal was to test instruction, norm, or rule compliance in textual context, and (iii) provided at least one quantitative metric. Thirty-three benchmarks met these criteria.

Each paper was then coded along four dimensions: the kind of constraint under test (instruction, norm, or rule); the form in which contextual information was supplied (from a static prompt through a multi-turn dialogue to a simulated web task); the verification method employed (scripted checker, symbolic-logic tool, LLM judge, human panel, or a hybrid of these); and the degree to which the benchmark engineered collisions among competing constraints. Because the corpus varies widely in purpose and granularity, the chapter offers a narrative rather than tabular synthesis, contrasting emblematic datasets, highlighting methodological innovations, and tracing recurring blind spots. Three structural gaps emerge — weak context representation, scarce collision testing, and limited manipulation of ‘normality’ assumptions — which motivate the interference-aware metric proposals developed in the dissertation’s final discussion.

3.2 Bringing the Cases Together: A Triangulated Comparative Strategy

Together these lenses translate the dissertation’s theoretical wager — that technical levers cannot be disentangled from accountability infrastructures — into an empirical programme capable of following harms as they migrate across code, policy, and lived experience.

The dissertation adopts a *comparative multi-case* design in which the three focal artefacts — synthetic-data pipelines, machine-unlearning procedures, and contextual-evaluation benchmarks — were selected as *critical cases*. Each occupies a distinct locus in the machine learning workflow: data acquisition, model maintenance, and system-level evaluation. Together they offer maximum variation along the pipeline axis while holding one analytic constant: all three embed the same underlying *individual-control* premise. By examining how that premise is enacted, strained, or bypassed at each stage, the study can expose structural limits that a single, more homogeneous corpus of evidence would leave invisible.

This comparative frame also explains the deliberate plurality of empirical approaches. Chapter 4 analyses a synthetic-data deployment through a blend of dataset forensics and regulatory document review; Chapter 5 uses a scenario-based elicitation to trace how data subjects weigh deletion, opt-out, and unlearning in context; Chapter 6 conducts a systematic survey of state-of-the-art evaluation benchmarks to map where instruction metrics collide with contextual norms. Treated separately, each method illuminates local dynamics; but read side-by-side, the findings crystallise into a set of *homologous failure modes* — lineage loss, relational trade-offs, and metric drift — that recur whenever individual-control logics confront the collective nature of sociotechnical harm. Methodological diversity thus functions not as patchwork but as triangulation, allowing the dissertation to trace a single analytic thread across otherwise disjoint technical domains.

Finally, the combination of dataset audit, interpretive elicitation, and benchmark meta-analysis constitutes a template for *multi-level socio-technical inquiry*. Recent calls for an ‘evaluation science’ of AI [272] emphasise the need to blend qualitative reflexivity with quantitative stress-testing. By stacking micro-level audits, meso-level user studies, and macro-level metric reviews within the same comparative frame, this dissertation operationalises that agenda and demonstrates how cross-method synthesis can reveal failure-patterns that remain opaque to any single lens.

Chapter 4

Case I: Risks of Synthetic Data

This chapter is based on the paper “Real Problems with Fake Data”, which received a Best Paper award at FAccT ’24, and was written in happy collaboration with Justin Norman, also of the I School.

4.1 Context

Why Start with Synthetic Data?

The dissertation’s first empirical site (Chapter 4) is *synthetic data* because it occupies the bulls-eye of the individual-control paradigm. In white papers and compliance decks, synthetic generators are heralded as a ‘privacy-by-design’ silver bullet: swap every identifiable record for an artificial twin and the notice-and-consent headache evaporates [152]. Regulators have begun to echo the promise. UK ICO guidance treats adequately de-identified synthetic data as non-personal, while U.S. health-tech firms market GAN-derived patient records as HIPAA-exempt evidence of responsible innovation. In short, synthetic pipelines translate the *legal fiction* of privacy self-management into a *technical mechanism* — precisely the intersection mapped in Chapter 2’s exploration of related work.

Empirical cracks in the substitution logic

Mounting evidence shows the promise is brittle. GANs and diffusion models memorise and regurgitate training snippets, re-introducing the very leakage they claim to eliminate [51]. Even when memorisation is rare, distributional fidelity carries a cost: synthetic faces skew lighter, younger, and more Euro-centric than their sources, boosting statistical utility scores while widening representational gaps [125]. Audits across vision and tabular domains find that fairness metrics improve in-sample but degrade on out-of-distribution groups once ‘balanced’ synthetic data is injected [15]. These studies make concrete the argument developed

in Section 2.3: a privacy fix engineered for *rows* can silently magnify harms embedded in the *columns*.

Lineage opacity and governance friction.

Synthetic data also strains enforcement infrastructure. Once transformed, a record’s provenance and licensing terms become, for practical purposes, lost, undermining FTC “affected work-product” deletions and frustrating GDPR erasure audits. Dataset curators, rather than data subjects or oversight bodies, decide which demographic cues are preserved, perturbed, or invented, consolidating epistemic power in developers’ hands [31]. Hence synthetic data simultaneously *strengthens* per-subject privacy claims and *weakens* collective accountability — the tension that animates both of this dissertation’s research questions.

Chapter 4 as a test-bed for RQ1.

The chapter therefore asks: *When synthetic data stands in for the real thing, whose privacy is protected, whose representation is distorted, and who loses recourse?* By combining statistical audit and regulatory analysis, we show how lineage-obscuring convenience becomes a vehicle for ‘diversity-washing’ and for consent circumvention. Synthetic data thus provides the clearest proof-of-concept for the dissertation’s central claim: solutions that perfect individual control often defer — or deepen — relational harm.

Foreshadowing the remainder of the dissertation.

The empirical arc that begins here proceeds in two steps. Chapter 5 examines *machine unlearning*, a tool celebrated for extending individual erasure rights but subject to similar relational blind spots. Chapter 6 turns the lens on evaluation culture itself, showing how benchmarks that measure instruction-following and contextual safety entrench a per-user gaze. By starting with synthetic data, arguably the paradigm’s most seductive “fix”, we ground the analytic strategy that carries through the dissertation: trace a privacy enhancement from its per-record design objective to its system-level externalities, then assess how laws, metrics, and organisational routines either surface or suppress those externalities. The chapters that follow replicate that method across distinct technical artefacts, cumulatively mapping the limits of the individual-control model and pointing toward relationally oriented alternatives.

4.2 Introduction

Facial recognition technology (FRT) has become commonplace, used from flight check-in at airports to police crowd-monitoring. Bias in FRT models has resulted in mis-identification and expanded surveillance, causing unjust incarceration and other discriminatory outcomes.

Attempts to solve these issues by increasing the accuracy of FRT run head first into problems; for a machine learning-based computer vision system to be considered robust enough for a given real-world task, it must ‘generalize’ to images that vary widely in quality and domain (image granularity, race, age, gender, background, head pose, hats, glasses, etc.). Datasets with this level of granular design and annotation, that are also large enough for use in deep learning, are nearly impossible to find due to logistical and ethical concerns. As a result, researchers have turned to synthetic data generation, where data is generated to resemble something without being a representation of an instance of it — a drawing of a generic face as compared to a photograph of a real person. Synthetic data has been used to augment existing datasets and create new datasets for better training and evaluation of FRT models. Logistical and ethical challenges to data collection exist outside of FRT, and synthetic data usage has become commonplace across machine learning, from computer vision to large language models. This paper examines two key risks of using synthetic data.

Synthetic data is fundamentally useful where real data is not fit to task, necessitating that synthetic data must be both similar enough to be meaningful, but different enough to mitigate the reason the real data is not usable [152]. Jordan et al. propose three attributes of synthetic data that must be met for it to function in lieu of real data: utility, fidelity, and privacy. This paper focuses on facial recognition because it clearly articulates the risks of synthetic data, inherently forcing trade-offs between these attributes. There is high difficulty in making a picture of a face private but still usable as training data (privacy vs. utility) — a face which has been obscured to the point where an identity could not be gleaned is less useful [52]. Achieving fidelity in facial datasets, a measure of how well synthetic data matches the real world, is also saliently difficult in facial recognition use cases, as we examine below.

The first risk we focus on is the *high risk of false confidence* in the ability of synthetic datasets to mitigate bias in data distribution and representation. We demonstrate this through the real-world example of using synthetic data for a facial recognition model evaluation. This paper was motivated by the realization of the under-explored risks of synthetic data while conducting the evaluation, and we present it both to provide an example of how synthetic data is used and to detail the concerns that conducting it made apparent. In brief, synthetic data offers a way of diversifying datasets, but diversity in real-world faces often follows from cultural practices that are qualitative and meaning-laden rather than quantitative. Creating a synthetic dataset or adding synthetic data to existing datasets in an attempt to diversify that dataset runs the risk of *diversity-washing* — appearing to resolve valid criticism regarding a dataset’s distribution and representation but in a way that is superficial. As a result, using synthetic data risks legitimizing technologies such as FRT despite potentially continuing to propagate bias by achieving false fidelity.

The second risk we examine is how using synthetic data risks *circumventing consent for data usage*, illustrating the impacts by considering the importance of consent to the U.S. Federal Trade Commission’s regulation of data collection and affected models. Synthetic data provides an avenue for model developers to side-step thorny issues around collecting large-scale representative facial datasets. Proper consent to data usage is foundational to

the privacy enforcement tools that the FTC has used to require companies delete ML models trained on improperly collected data, a key regulatory hurdle to improper data collection and resulting harmful model deployment. Using synthetic data risks circumventing and obfuscating consent, thus complicating deterrence and enforcement.

This paper proceeds as follows: We begin in Section 4.3 by summarizing related prior work. We first examine work on datasets and representation, before discussing participation and consent and power over data and models. Finally, we discuss synthetic data and its use. We then proceed to the two titular risks of synthetic data that this paper focuses on — diversity-washing (Section 4.5) and the circumvention of consent (Section 4.6). We draw upon two real-world examples: a facial recognition evaluation task conducted using synthetic data, and the FTC’s enforcement actions against models trained on deceptively collected data to illustrate these risks. Finally, in Section 4.7 we expand upon how these two risks are examples of irresponsible use of synthetic data: consolidating power in the hands of model creators, and decoupling data from those it represents and those who are harmed by its improper use. It is our intention for this research to contribute to the field by presenting tangible examples and background for the challenges inherent in responsible use of synthetic data, thus laying foundations for further work and debate. We call for future work to examine the breadth and usage of synthetic data and to work towards both mitigating synthetic data’s risks and enabling its potential for participatory empowerment.

These twin risks — diversity-wash and consent circumvention — speak directly to the dissertation’s research questions. Both originate in the tool’s attempt to secure privacy at the record level while ignoring relational context. In what follows we document those risks empirically and analyse how they exploit the blind spots of notice-and-consent governance.

4.3 Related Work

We review literature along three axes that mirror the individual-control → relational-harm → governance triad. In this section, first, we focus on the datasets that underpin machine learning systems, and detail how that work treats the specific issue of data distribution and representation in those datasets (Section 4.3). Next, we discuss prior work on participatory governance, consent and data privacy, and attempts to capture some power over dataset creation and usage by those most affected (Section 4.3). Finally, we provide a summary of work detailing what synthetic data is and how it is used (Section 4.3).

Datasets, Diversity and Representation

Datasets are the hidden infrastructure behind machine learning, most visible when the systems built on them break [147]. Models developed and dependent on these large datasets can lead to biased and harmful effects, with models used for bureaucratic categorization in particular having a long history of harm [43, 9, 249]. The collection of data is then frequently the starting point for ML-disseminated discrimination and bias in domains such as

hiring [216], advertising [164], pricing [281], the application of law, and government allocation of resources [3]; as well as being vital for identifying and enforcing against discrimination [10]. The stakes of responsible dataset development, then, are high, and we build on critical previous work [205, 208, 140] in focusing on the ways that dataset creators have significant impact on the harms that occur downstream via their development, usage and deployment [156]. More narrowly, we hope to bring focus to important risks present when synthetic data is used in the process of creating and using datasets in machine learning development.

Datasets used for facial recognition models, where the goal of the model is matching an image or video of a face to an identity, have received much scrutiny — specifically for violating privacy [132, 32]. An analysis by Crawford & Paglen [69] of ImageNet, a frequently-used large dataset, demonstrated active labeling of faces with offensive and derogatory classifications, and Prabhu & Birhane [32] make the point that beyond obvious privacy harms such as blackmail, the creation of one of these datasets causes similar datasets to propagate. The recent identification of CSAM material in the popular LAION dataset is an example of this [257]. In the case of synthetic data, where the data is frequently derivative of previously collected data (as expanded upon below in Section 4.3), this then risks the continued propagation of non-consensual imagery. The above highlights the need to focus beyond just the models. Much of the critical AI literature focuses on vital interventions to change model outputs that discriminate against protected classes. This work, instead, more closely follows work such as Buolamwini & Gebru’s “Gender Shades” [44] that is focused on the *data* which is fundamental to AI system development.

To understand the risks posed by using synthetic data to create and add to datasets used in machine learning development, it is necessary to understand the landscape of both machine learning development and the stakeholders impacted by it. We will use the taxonomy of dataset development stages and subjects presented by Khan & Hanna [156]. We lean on this taxonomy throughout the paper, finding it to be clear in its intention of “providing a common language for conversations across datasets” between practitioners, scholars and regulators. Starting from first principles, machine learning is not rules-based like traditional software development, but instead consists of a practitioner teaching a model to identify patterns in a dataset. To begin, Khan & Hanna assert, the *task* for which the model is being trained must be formulated and constrained. Next, *data* must be collected, meeting the constraints of what is necessary to train a model to achieve said task. That data collection is usually broad, requiring the data be *cleaned* before it is annotated. The *labels* attributed to data by the annotator are of vital importance to machine learning systems, as the systems are taught to identify those labels in their training data. After this point, model *training*, *valuation*, *implementation*, etc. may occur. Khan & Hannah define multiple *stakeholders* in the process of creating the datasets used for ML development: the curator, the data annotator, the data subject, the copyright holder, and the model subject. The *curator* is the entity responsible for dataset creation, while the *annotator* is (frequently outsourced [120]) responsible for annotating the dataset. The *data subject* is the person whose biometric information is present within the collected data, the *copyright holder* may hold exclusive

rights over that data¹, and the *model subject* is the person who is impacted by the decisions made by the model trained on the data. The last three categories are fluid, and can consist of the same person or of two or three distinct entities.

Creating a training dataset that is representative of model subjects is challenging, and when done poorly, results in inaccurate and frequently harmful outputs; here we consider how such harm might arise across different stages. In producing a representative dataset, different dimensions of identities are frequently missed. Datasets are frequently biased by necessity to meet a specific intended use of a model, but when categories are socially constructed (such as race and gender), the observed, inferred data that dataset curators use to bound and constrain data collection (and that annotators must use to label) can clash with how the data subject self-identifies. Biased data representation is also a concern, with annotation reflecting social biases and stereotypes across gender, race, and more [229]. This can lead to rampant misrepresentation and miscategorization of both data and model subjects, producing forms of control. Annotation work is inherently an interpretive project, but results in data that is perceived to be ground truth [39]. In reality, as shown by Recht et al. in work on testing the generalizability of ImageNet, when attempts to replicate annotation are made, different distributional properties for the same data emerge [222]. Annotation is also frequently the cause of artifacts in datasets that allow for models to overfit to training data when solving a task, with many of the concerns arising from how human data annotators are instructed to label [285]. Beyond concerns with annotation, the question of what data to collect when constructing a dataset to correctly answer a question is complex. The task of annotation itself presupposes that the question being asked of a model is one that can be answered — Aguera y Arcas et al. demonstrate how a model trained on ‘gaydar’ data was in reality labeled around stereotyped aesthetic traits, showcasing an example of labels being generated not because of any model-pertinent aspects of the data, but rather simply because the question was being asked [11].

Beyond the above challenges to creating a representative dataset, there are ethical issues that are raised by efforts to produce a dataset that represents a diverse community. Data collection requires infrastructure, and that infrastructure is frequently co-constitutive with surveillance infrastructure. Even when data collection is initiated in service of providing services to the most disenfranchised, rendering the members of those communities hyper-visible frequently serves to hurt those same communities, as decisions are made for them by others [10]. These decisions can reinforce oppressive norms, such as visual gender binaries [34, 129], further delegitimizing disenfranchised groups in a clear example of administrative violence [249]. Even when categorization schema of data subjects are correct, their use as prescriptive instead of explanatory can lead to attribution errors, co-opting classification in a oppressed group as a reason for that very oppression. Machine learning systems used to predict recidivism are a prime example [58], where factors like race, which make a group member more likely to be targeted for discrimination, are frequently used instead as

¹Not all data is copyright protected, and even when it is, different legal regimes have different limitations and exceptions to exclusive intellectual property rights.

a predictive factor when individuals are made model subjects.

Connection to dissertation: Representation harms originate in the dataset stage — harms that notice-consent remedies leave untouched and that Chapter 4 will surface in synthetic-data form.

Participation, Consent and Privacy

Participatory approaches are frequently fronted as a way of mitigating the harm that results from AI systems, both at the dataset level as described above, and in model training and deployment. These approaches focus on engaging the public, and build on policy approaches such as feedback sessions, public hearings and impact assessments [114, 138]. Participatory design methods in particular focus on co-design to incorporate user context, needs and values [41, 242, 146, 128], designing systems *with* those affected instead of *for* them. Recently, participatory AI work has explicitly focused on those for which AI most frequently exacerbates harm [224, 145]. Patel et al. [204] draw from previous work, including Arnstein’s influential Ladder of Citizen Participation [13], to detail five levels of participation in data stewardship, including: 1) *informing* people about how their data is used through methods such as model cards, 2) *consulting* people through UX research and surveys, 3) *involving* people in data governance through panels and public deliberation, 4) *co-design* of data governance and consequent technologies through structures such as data trusts, and 5) enabling *decision-making* through citizen-led governance boards. We will return to Patel’s framework when discussing the risk of synthetic data circumventing consent (Section 4.6).

There is a wide — and growing! — diversity of participatory work in AI. Examples range from crowdsourcing impacts [21, 76] and data labelling [201] to eliciting preferences for dataset collecting and design decisions [57]. Peng et al. [208] recommend that dataset creators make ethically salient information clear and accessible while actively stewarding the dataset and its future use, and encourage retrospective study of datasets due to the difficulties in understanding issues at the beginning. Hutchinson et al. [140] detail documentation requirements at each stage of the dataset development lifecycle, with different document types for each stage, and call for frameworks for transparency and accountability. These fall across the range of Patel et al.’s framework, [204] and critiques of these methods include characterization of it as ‘participation-washing’ [113, 243], with Arnstein describing approaches such as public requests for comment as “tokenizing” and “inadequate in shifting power” [13]. Sloane et al. [243] argue that these approaches can function as unrecognized labour, and the line between tokenization and participation in cases such as crowdsourcing is quite blurry. Birhane et al. [33] show examples of community inclusion in annotating datasets, improving documentation and increasing the utility of large language models for under-served languages, and other examples include community organizations such as the Detroit URC3 which evaluates potential partnerships between community organizations and researchers to avoid exploitation [66], and examples from Indigenous Data Sovereignty [217]. At a large scale however, there are still major hurdles. Groves et al. investigate the hurdle of making participatory approaches work in the commercial AI labs that are the primary site

for AI research, and find that “corporate profit motive and concern around exploitation are at present functioning as significant barriers to the use of participatory methods in AI” [124, p. 10].

While the above participatory approaches centre shifting decision-making power to include data subjects and model subjects, these approaches tend to require model creators to opt in at least for now. Consent, though enacting a more limited form of participation, *requires* model creators to be wary of unfair and deceptive practices that overstep expressly-informed consent when collecting and using data. Indeed, consent violation is a legally cognizable privacy harm, one with potential repercussions. In the U.S., state information privacy laws do some work to enforce this, with the Illinois Biometric Privacy Act (BIPA) both resulting in a significant number of lawsuits alleging violation, and being responsible for the largest settlement amounts from companies who have breached BIPA by deploying FRT [251, 287]). For instance, in *Vance v. IBM*, the court affirmed that IBM violated BIPA by not receiving written consent before collecting and disseminating individuals’ images in their “Diversity in Faces” dataset [115], even though the images used were public. Publicly accessible personal information comes with an intended context of use, which can be violated by memorized and regurgitated data [51]. As will be discussed in Section 4.6, to date, the Federal Trade Commission’s enforcement power around unfair and deceptive data practices has centred upon the absence of consent. Practically, this instils consent as the most direct way for data subjects and model subjects to participate in decision-making around the models which affect them, albeit mostly *ex post facto* through their ability to prompt enforcement when discovering that their consent has been violated.

Consent violations frequently occur through improperly scoped consent, where data collected for one purpose is repurposed. This can result in adverse effects beyond the concrete privacy harms [247] that are most often legally enforced in cases such as data breach, e.g., identity theft. As an example, data used beyond its consented purpose leaves data subjects at risk of discrimination harms, facing miscategorization and expansion of surveillance, as detailed above. Such data can also be sold and shared with third parties, further increasing the odds that it is not being used for the purpose it was collected, and therefore that it is frequently in violation of the consent of data subjects [48]. Even when consent is nominally obtained, transparency is often in name only, with data subjects overwhelmed by opaque and all-encompassing digital policies, terms, and conditions [203].

Ultimately, the question of consent is complex. Brown et al. [42] argue that the current paradigm of training on publicly accessible data makes it highly challenging to distinguish what public data was made public with blanket versus contextual consent, and that, therefore, obtaining informed consent is difficult at best. They make the case for training solely on data explicitly consented for public dissemination. We will return to this argument in Section 4.7, as it supports using synthetic data generated from properly-consented real data or responsibly procedurally created data, and criticizes using synthetic data generated and used in a manner that exacerbates concerns around consent.

Connection to dissertation: Consent frameworks secure individual control yet falter when data lineage blurs; Chapter 4 shows how synthetic data exploits this blind spot.

Synthetic Data

Synthetic data in machine learning is defined by its driving goal of mimicking real-world data — it is synthesized to be used as though it were real data for training machine learning algorithms [149]. It differs from what is usually referred to as ‘data’, i.e. non-synthetic data, in that it does not have an explicit 1:1 real-world referent. When training a computer vision model to recognize a face, the data traditionally used are representations of real faces, photographs taken of real people. The same holds true for other forms of data — scientific data records representations of physical phenomena such as sensor readings, natural language data is text composed by a real person, etc. Synthetic data is made to resemble these things, but is not explicitly a representation of a real thing. Using the example of a face, a synthetic face could be a drawing that looks for all intents and purposes like a face, but that is does not represent a specific person.

In actually creating synthetic data, however, things become muddier. The term encompasses data generated by generative models, more traditionally augmented data, and procedurally created data [171]. We differentiate between the first two and the latter category based on how derivative of a real-world training dataset they are. *Generated data* is the output of generative models: ML systems that produce a (supposedly novel [51]) output from an input by abstracting over their training data. One generative model that has captured popular attention [209] is StableDiffusion, which generates art from a user-provided input sentence [49]. *Augmented data* is fuzzier, but equally derivative of a real-world training dataset; the term tends to refer to any real-world data to which modifications have been made. A model creator seeking to increase the performance of a model on its specific task may create many versions of each image sampled from the input dataset, creating augmented data. This type of synthetic data cannot be considered inherently private or unbiased, with generative models explicitly being found to frequently regurgitate memorized training data [17]. At larger scales, this type of data can be used as training and evaluation datasets too. As detailed by Khan & Hanna [156], datasets are vital components the larger model development cycle, priming synthetic data to reinforce and scale skewed values and requirements that are embedded within models, datasets, and benchmarks.

Generated and augmented data differs from techniques for *procedural creation* of data, where dataset designers make active decisions to create ‘net-new’ representations of data similar to what might be found in the natural world. There are some important differences. The easiest way to visualize how procedural creation works is considering video game character creation — a player starts with a base face shape, and adds the features they want. Notably, with procedural creation of faces, the base volumetric face scans that are used are very far removed from the people that were scanned — they are low fidelity representations, making data lineage even murkier. Instead of outputs being generated from, and therefore bound by training data, procedural creation can result in ‘net-new’ data that never existed previously.

Or, as an example outside of computer vision and facial recognition, consider procedurally created finance data which uses agent-based models to mimic real world data generation by

creating representative agents and attempting to model money laundering [175]. It must be remembered, however, that both agent- and procedural-based synthetic data are highly determined by preconfiguration and design. As such, making inferences about the real world based on procedurally created data is difficult at best. Recalling Jordan et al.’s framework of synthetic data’s usability, this type of synthetic data faces utility and fidelity hurdles [152].

Synthetic data has frequently been explored as a method to avoid privacy concerns, increase model performance and to reduce model bias. Privacy concerns have the longest history of motivating synthetic data [152]. Healthcare [116] and financial [14] domains have been particularly attracted to synthetic solutions due to the sensitivity of their data. Examples including simulation studies in population health [193]; synthetic clinical records used for IT development, education, and training [72]; money-laundering detection [175]; and public release of augmented financial and healthcare datasets to enable open science and research [131]. In contexts of societal bias, synthetic data has been explored as a way to remove disparate impact [100, 153, 289], to suppress imbalance effects and to racially balance datasets [161], as well as to remove sensitive information and blind models to race [268]. However, the latter has been found to not always be effective in practice, with applying a ‘veil of ignorance’ not having any notable influence on accuracy of FRT on under-represented categories [271].

Increasing model performance by using synthetic data has usually meant enlarging datasets to provide robustness to outliers [278, 94, 70]. Additionally, operating at a slightly different scale of enlargement (from 0), it has been used in situations where real world data is difficult to access. Google recently demonstrated AlphaGeometry, an AI system purported to solve “Olympiad geometry problems at a level approaching a human-gold medalist”, trained solely on a dataset of 100 million synthetic math proofs — a dataset which could not exist using human generated proofs [260]. Computer vision systems require (often impossibly) large amounts of labeled training data in a specific domain [32].

Finally, consider the context of facial recognition. Though large, open datasets specifically developed for facial recognition tasks exist, such datasets are either extremely basic (i.e. passport photos with great lighting), too narrow (only contain a biased subset of race, gender, head/body pose, etc.), or simply contain glaring and challenging shortcomings [218]. As a result, dataset creators utilize synthetic data. In FRT, this is either (1) procedurally created ‘synthetic’ data, where a bone structure scan is used as a basis for volumetric face models, and then textures representing features are stretched across that model and swapped out [288], or (2) the perturbation of existing data to produce more diverse samples from the existing distribution, including both simple techniques and advanced techniques such as diffusion models and generative adversarial networks (GANs) [75]. We expand upon the use of synthetic data for FRT in the following section.

Connection to dissertation: Synthetic generators operationalise per-record privacy but intensify relational bias and enforcement opacity — the dual risks empirically examined in Chapter 4 and revisited throughout the dissertation.

4.4 Methods

The chapter examines two failure modes of synthetic data — *diversity-washing* and *consent circumvention* — by retracing a real face-recognition evaluation and by reading publicly filed U.S. Federal Trade Commission (FTC) orders. The first strand is technical, the second doctrinal, but both address the same question: how does ‘privacy-by-substitution’ behave once it leaves the white paper?

Technical strand: evaluating face recognition with synthetic data

Base corpus.

The starting point is **CASIA-WebFace**, a celebrity set long used to train and benchmark open face-recognition systems. Its demographic skew — lighter skin tones, Western media personalities — is well documented and therefore auditable.

Partially-synthetic probe set.

To mimic real-world CCTV conditions the authors construct a ‘digital line-up’ protocol. For each identity a sharp source image is chosen and embedded; the ten nearest neighbours in CASIA form a candidate line-up. The source image is then *systematically degraded* along three axes (blur, resolution down-sampling/upsampling, JPEG compression) in three steps each, producing nine quality tiers. The degraded image is the probe, the original gallery image plus its neighbours form the line-up, and the task is to return the correct identity. Because the degradation is applied after sampling, demographic composition mirrors CASIA: the corpus is ‘partially synthetic’ but still derivative of real faces.

Fully synthetic gallery.

A second corpus is generated with the commercial *Synthesis AI* platform. The API randomises age, coarse gender, pose, illumination, expression, hair and background, but the underlying 3-D head meshes come from a proprietary library of 511 base scans. When the authors request ‘balanced’ output, the generator nevertheless returns a gallery dominated by White male faces and almost no Asian, Middle-Eastern, or Black women, illustrating the ease with which statistical ‘balance’ can mask representational gaps.

Evaluation pipeline.

A publicly available ArcFace-style ResNet-50, trained on unaltered CASIA-WebFace, serves as the reference recogniser. Performance is recorded as the fraction of probes whose top-1 match is the correct identity at each degradation tier.

Doctrinal strand: consent circumvention in U.S. enforcement

To see how lineage opacity frustrates privacy self-management, the study reads every public FTC consent order that mandates *deletion* of a model trained on data collected without valid consent: *Everalbum* (2021), *Clearview* (2022), *Ring* (2023), *Rite Aid* (2023), and *Kurbo* (2022). Because the FTC files are public, the analysis is purely documentary; no interviews or confidential discovery material are used.

Limitations

The technical audit is confined to face recognition, to a single open training set, and to one commercial generator; other modalities may expose different trade-offs. The policy reading is limited to what appears in the published consent agreements; any sealed evidence remains out of view.

By pairing an industrially plausible evaluation with a close reading of enforcement doctrine, the chapter shows that a tool calibrated to protect *individual rows* can still distort group representation and erode the consent-based levers that regulators rely on — an empirical foundation for the dissertation’s broader critique of individual-control governance.

4.5 Risk 1: Diversity-Washing

This section presents an example of using synthetic data; first describing a dataset that is partially-synthetic, a blend of augmented data and real data, and then describing a dataset that is procedurally created synthetic data. In both cases, we detail how using synthetic data risks creating datasets — and subsequently training and evaluating models from that data — that fail to mitigate bias in data distribution and representation. Furthermore, there is a risk of propagating harm through a patina of legitimacy, and through diversity-washing potentially harmful models.

The example we use is a real world example, where one of the authors had previously created synthetic datasets to evaluate facial recognition technology (FRT) [198]. We present this example to illustrate a risk of synthetic data and ground it in a real world setting — this is not an attempt to present novel work on FRT evaluation.

To provide a brief background: FRT is created with the aim of matching images of identifying faces. Companies that sell facial recognition such as Clearview often tout accuracy rates of their systems of 97% or more [254] — but these calculations are often made under ideal conditions. In real world conditions, such as surveillance camera footage, captured images of faces may be poor quality. FRT has been shown to be prone to make erroneous matches (i.e. identifying someone incorrectly as a match) when using low quality images as input [136]. However, users of these systems, such as the police, and adjudicators such as judges or members of Congress, who are not experts in FRT or ML/AI, are at a distinct disadvantage in evaluating companies’ claims.

Augmented Partially-Synthetic Dataset

In order to evaluate FRT models in real world settings, first, benchmark performance for FRT models needed to be established. This occurred by stimulating the (highly unreliable) process of a human identifying an individual from a visual line-up of other humans with similar characteristics. To do so, a source image of a selected identity was identified from the base dataset, detailed below, and a ‘digital line-up’ of (mathematically) similar faces from that base dataset were created. Augmented data was then created by progressively degrading the image of the source identity, and then this augmented data was compared to the similar identities in the digital line-up, as well as to the source, in order to mimic real world settings. The success of the evaluated models was defined by the rate the correct identity was selected with the augmented data, the degraded source images, as input.

A significant body of knowledge already exists concerning both the obvious and non-obvious potential harms in gathering image data containing human subjects, and the real harms of processing such information through FRT [218]. As such, it was important to begin with core datasets that had already been evaluated thoroughly in the literature, rather than collect wholly new human subject data. As such, CASIA-Webface, one of the two most widely used and evaluated public datasets [288, 155], was selected as the dataset for use as both non-augmented data and as the base for augmented (in this case, degraded) data. This was chosen due to its sourcing from crawled and scraped publicly available images of celebrities, strict rules prohibiting commercial use, wide variation in image quality, large number of identities (depth), and large number of images per identity (width) — important features for the dataset that was the provenance for later augmentation. As detailed in Section 4.3, this choice of base dataset is inherently political despite frequently being rendered neutral. Using it as the base for the creation of synthetic data makes it inherently more so due to the downstream effects of using the base dataset in generating derivative data.

The data augmentation techniques used to generate the augmented portion of this mixed dataset can differentially and unpredictably scale issues — making images black and white, as an example, could further segment training data by racial presentation. Model creators frequently augment data while training models, past the stages of development where they are considering dataset collection and annotation. This process has a set of well documented risks for model fairness. However, in using this process to create large base datasets, there is a change of framing that re-introduces these risks. For example, the dataset created for the FRT evaluation (created by augmenting data and combining with real data) was created with the explicit goal of being more representative of real world conditions. Datasets are frequently treated as ground truth [32], hiding the decisions and processes by which they were created. This risks ignoring issues that can occur from augmenting data. Even if synthetic data appears ‘diverse’, the generation of that data cannot be unwound from the particular datasets and models that it is being generated from, and any attendant shortcomings. To start, any biased representations would at best replicate from the original dataset. If data augmentation technique(s) impacted some subjects differently than others, the resulting impact could be unintended bias in the dataset. Since the presence of such relationships

are rarely known, much less understood statistically in datasets, it is also possible that the sampling strategy used to choose which data from the dataset will be used in training may actually exacerbate harm by over-representing biased representations. Deep learning techniques are generally already susceptible to overfitting, where a model learns how to predict patterns in a way that pays too much attention to the training data, and doesn't generalize to other data — it learns specific idiosyncrasies and meaningless data artifacts. Synthetic data seems like it should have the capacity to remedy overfitting, through careful and bespoke dataset construction that debiases data distributions. One could assume this would increase fidelity and enable better generalization over a more diverse training space. In reality however, when synthetic data is overfitted, these idiosyncrasies can go through the entire model training process unnoticed. As such, synthetic data instead increases the likelihood of overfitting errors being propagated through, necessitating that further technical care is taken to prevent overfitting. In our FRT evaluation example, such preventative measures were taken by curating the dataset so that visible artifacts such as skin tone were at parity with acceptable real world dataset distributions. However, the perils of overfitting are a way in which synthetic data can struggle to meet the standard of utility necessary to work as a replacement for real world data [152].

Procedurally Created Fully Synthetic Dataset

Beyond the above partially-synthetic dataset, there was a need to better evaluate performance on specific types of data not present in the original datasets that the FRT models were trained on. So, a synthetic dataset consisting of procedurally created data, namely mixed examples of non-degraded and degraded computer-generated faces [213, 25], was developed. As previously detailed, this is a common usage of synthetic data — needed representative data was not available for collection, and so generating synthetic data was the easiest method of proceeding.

The Synthesis.AI software² used to create the fully synthetic dataset (as with most procedural synthetic human/object generation tools) works by providing unprecedented control over how a dataset and its inherent metadata parameters are specified. This software employs a combination of classic rendering and generative synthesis to create photo-realistic images of human faces, bodies, and environments [195]. A user is able to decide how much and which type of each characteristic (in our case age, race, gender, hair type, pose, lighting etc.). However, the tool did not make any suggestions or place any controls based on sociotechnical norms or demographic data (such as the census etc.) when creating a synthetic human dataset of any type. When first testing the Synthesis.AI API, a dramatically racially imbalanced dataset was returned, even though the specification given was for randomization of the race characteristic. At first glance, the dataset appeared diverse and was numerically at parity for gender. However, the software lacked permutations for Asian people, Middle Eastern people and Black women, leading to a stark racial disparity upon deeper inspection,

²Synthesis AI (<https://synthesis.ai>)

and a preponderance of white men and white women despite attempts at balancing racial demographics. Such a system allows any user to easily create an unintentionally biased dataset, which could then be used to train a biased model. Instead of mitigating data distribution and representation concerns, this risks extending them.

As a further example, Microsoft’s FaceSynthetics [279] is a procedurally created synthetic dataset of 100,000 individuals, with faces derived from representative 511 base scans. However, these 511 base scans include only 30 Black men, and even fewer Hispanic/Arab/Indian men and Black/Hispanic/Arab/Indian women (borrowing the reported demographic categories), meaning that the fully diverse population they claim include multiple racial categories fully defined by the ways in which these 30 faces can be manipulated through a generative process. These manipulations include fine-tuning hair, expression, and clothing, but published details on the process of how these potentially racially-coded aspects were chosen are sparse. It is not known how those features are distributed in real faces, and attempting to extrapolate a portion of a representatively diverse dataset from such a small set of base faces leads to a risk of *statistical diversity without representational diversity*, compared to a representative dataset of real images with both statistical and representational diversity. Synthetic data here falls flat in addressing these complex, cultural and deeply contextualized factors.

These tools risk falling into the ‘panacea of legitimization’ that Frank Pasquale describes [203], where ethical concerns are not only routed around, but are routed around in such a manner that they can reify malpractice due to the co-constitutive nature of ML practices and computing platforms [30]. We point to recent work focused on toolkits for supporting practitioners in contextualizing ML system work as an avenue for improving upon this [73].

4.6 Risk 2: Circumvented Consent

Consent has become a key component of privacy and data protection, both via regulatory enforcement and as a necessary foundation for the participatory approaches that have emerged as ethical practice for preventing harm. Consent also plays a role in U.S. sectoral statutes such as HIPPA, U.S. state laws such as California’s California Privacy Rights Act and Illinois’ Biometric Information Privacy Act, as well as laws in many other countries, with the EU’s GDPR a notable example. These statutes share a common goal: to provide people with control over their personal data, via notification, access, and consent regarding the collection, use, and disclosure of personal data. This type of privacy regulation is referred to as ‘privacy self-management’ by Solove [247], and focuses solely on whether or not data subjects have consented, rather than on value judgements of privacy practices. This section will focus on illustrating the risk that synthetic data poses to consent-based frameworks by expanding upon how the Federal Trade Commission (FTC) has functionally used consent as a key aspect of conducting enforcement against companies using ML systems. Because deception/consent is assessed record-by-record, synthetic data can mask violations without

touching relational bias. The analysis is guided by one author’s experience at the FTC, but draws upon solely public knowledge.

Consent, Deception and Model Deletion

The FTC plays a vital role in the current U.S. privacy legal framework. This framework emphasizes individuals’ notice of, and consent to, the collection and use of their data. The FTC is an independent agency of the United States government that is tasked with protecting consumers and promoting competition in the marketplace. In the absence of federal privacy law, the FTC has played the role of de facto privacy enforcement, primarily based on its authority to police unfair and deceptive business practices. The Federal Trade Commission Act, and specifically Section 5, is a broadly applicable federal statute prohibiting “unfair or deceptive acts and practices”³. An *unfair* practice “causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition”, while a *deceptive* practice includes “any ‘representation, omission, or practice’ that is (i) material, and (ii) likely to mislead consumers who are acting reasonably under the circumstances” [248]. Notably, deception does not require any proof of intent. The FTC has brought deception claims against companies who have violated the terms of their privacy policies, failed to uphold promises of data security, or have failed to provide sufficient notice regarding data collection and use [248].

In settling cases against companies that have deceptively collected data, the FTC has required not only that the data in question be deleted and the affected users be notified, but also that all “affected work product” [167] be deleted as well — including models trained on that data. This approach is referred to as *model deletion*. The FTC posits that this approach is necessary in order to ensure that companies do not profit from the unfair or deceptive collection of data, and to prevent them from using the data in the future. Intellectual property is typically a tech company’s most valuable asset; it is an important factor for securing venture capital funding and the sale or licensing of IP often comprises tech companies’ core business models. In forcing a company to delete models, the FTC has also significantly changed the deterrence calculus for companies: from paying (relatively) small fines, to potentially losing a vital business asset [85]. The FTC has used the concept of model deletion in recent enforcement actions against companies that have collected data deceptively, including Amazon Ring, Rite Aid, Everalbum, Clearview and Kurbo (WeightWatchers) [141].

Synthetic Data and FTC Enforcement

The use of synthetic data risks undermining the utility of deception-based enforcement in regulating the collection of data, and therefore also undermines the regulation of models trained on such synthetic data. As previously described in Section 4.3, datasets play a

³15 U.S.C. Sec. 45(a)

foundational role in the models trained on them, and trusting models trained on deceptively collected datasets to operate without harm seems foolhardy. Enforcement by the FTC has hinged upon arguments that data was collected and used deceptively — often argued due to the absence of proper consent. By using synthetic data, however, it becomes easy for model creators to obfuscate the origins and consent of the data being used to create models. In the case of a procedurally created synthetic dataset, consent is no longer a procedural hook to limit downstream harms flowing from use, while in other synthetic datasets, unless data lineage is carefully recorded, traceability to the original data is at risk [228, 74].

Synthetic data also exacerbates existing logistical challenges for model deletion as an enforcement tool. Synthetic data brings questions of data lineage to the forefront, as ever-more-complicated sets of original, augmented, and derivative data are produced based on new face datasets with millions of people. As a small example, keeping track of whether a single version of a dataset has undergone ethical testing, or was sectioned off as a test dataset, is a challenge for FTC enforcers — let alone when datasets include different scaling factors and different degradations, with different subsets of identities (generated, procedurally created or real) and different levels of augmentation. One of the key hurdles to model deletion is the requirement for a high level of internal company documentation and logging. This documentation and logging is essential to identify the data that was collected illegally or deceptively, as well as the work product that was developed using that data. However, not all companies have robust internal documentation and logging systems, which can make it difficult for the FTC to determine the extent of the harm caused by the illegal or deceptive data collection practices. Another challenge is authentication and audit. Companies must demonstrate that they have successfully deleted the affected work product, and the FTC or other enforcement bodies must have a way of verifying that the company is being honest. However, this can be difficult, as it requires a level of trust in the company and its processes. In a setting where the FTC has to this point relied on settlement agreements, synthetic data further complicates existing logistical challenges, presenting important friction to enforcement.

Beyond Deception

In considering enforcement against companies using ML systems, it is important to note that beyond deception, the FTC has also enforced its unfairness authority. This occurred in the case of *Rite Aid*, where a biased facial recognition model was used to falsely identify people in certain protected classes as more likely to commit crime. This case was first-in-kind, but demonstrates that the FTC is not solely beholden to deceptive data collection as an avenue for enforcement. However, the details of the case were particularly egregious, with Rite Aid failing to undertake even the most basic risk assessments, and in part hinged on a violation of a previous settlement. Additionally, the system was trained on in-store camera footage without consent from data subjects, and model subjects were not notified or able to opt-out. Thus if Rite Aid had been investigated for deceptive data collection, harm

resulting from this system could have been prevented at the point where the system was trained non-consensually. But what would happen if Rite Aid had trained its FRT model on synthetic data? FTC enforcement would have to hinge on unfair practice alone. While successful here, the case against Rite Aid was egregious. The Supreme Court’s neutering of the FTC’s power to levy fines [56], in addition to both the deception and unfairness enforcements occurring through settlement rather than being decided in court, means the boundaries of the FTC’s ability to engage in this type of enforcement are still undefined. As such, the FTC’s ability to intervene both at the data collection stage (deception) and the model deployment stage (unfairness) gives options⁴. Synthetic data complicates the usage of a demonstrably useful tool for protecting data subjects and model subjects, by complicating the use of the deception standard.

Finally, it cannot be forgotten that while risking increased friction and obfuscation, synthetic datasets composed of augmented or generated data are demonstrably derivative, inherently based on real data representing real data subjects. And, despite the veneer provided by language such as ‘net-new’, procedural creation of synthetic data is also derivative, and thus suffers from issues of consent and participation too. In the example discussed in Section 4.5, the procedurally created synthetic dataset for FRT evaluation, the dataset was generated using Synthesis.AI’s commercial software. Software tools such as Synthesis.AI often utilize face and body scanning technology as the foundation of their generative processes, raising concerns around the limits of informed consent. The data subjects upon which these technologies are trained are rendered invisible and thus the use of such software is predicated on, at best [219], ambiguous consent. Similarly, it is important to acknowledge that the data subjects whose likenesses are captured in the CASIA-Webface, while primarily scraped from public sources, were not asked for informed consent regarding their data’s use — even when it has been shown that having your face included in such a dataset increases the accuracy of facial recognition models on your specific face [80]. As detailed by Peng et al. [208], using derivatives of common datasets introduces scaling concerns around propagation of improperly consented data, and as such, using synthetic data risks *further* scaling propagation of this issue. In decoupling data subjects from their data, this also removes their capacity to participate. Reconsider the Participatory Data Stewardship Framework mentioned above in Section 4.3; all five levels require that data subjects have at the very least visibility, and preferably control, over their data [204]. In further removing the ability for data subjects to consent, not only is that minimal level of agency reduced, but the potential for involvement in decision-making that directly effects them is erased. This illustrates the governance gap that Chapter 5 will revisit when we study machine unlearning — a tool likewise celebrated for enabling erasure yet fraught with collective-impact blind spots.

⁴While there exists the potential for both unfairness at data collection and deception at model deployment, cases to-date have lined up in this fashion.

4.7 Discussion

From Individual Privacy to Relational Harm

The positioning of synthetic data as a panacea to problems of representation and deceptive data collection, furthered by its portrayal as synthetic, as neutral, as created without lineage, risks placing the means of fixing those problems in the hands of those who created them and trusting that they'll get it right. Instead, the kinds of racialized misrecognition and bias that Ruha Benjamin, Safiya Noble, and others have drawn attention to must be considered when determining whether to use synthetic data. As Benjamin argues, our current machine learning development ecosystem must reckon with a history of discriminatory design in which racist values and assumptions are built into our technical systems. The “new Jim Code”, as she terms it, works to deepen the production of disparate harm, even while cloaked in neutrality and the language of innovation [28, 144]. Discriminatory practices are inherent to the current state of AI system development, privileging whiteness and discriminating against people of color, specifically women of color [197]. Sara Ahmed's work on the phenomenology of whiteness highlights the danger of a solution that further enables a reification of non-whiteness as a space outside. She details that “institutional spaces are shaped by the proximity of some bodies and not others: white bodies gather, and cohere to form the edges of such spaces” [7]. Synthetic data as a fix in this racialized context risks further enabling amplification of racial hierarchies, allowing for those within the boundaries to actively constitute the exclusionary and weaponized edges of these spaces: to define a face, train a model based on that definition, and decide its performance based on labeling racial boundaries. It risks not alleviating but instead contributing to *race as a technology*, designed to “stratify and sanctify” social injustice in the architecture of everyday life [28]; an added consolidation of power.

Lineage, Power, and the Limits of Consent Governance

Another example of the risks of consolidation of power through synthetic data arises when considering the inherent relational aspects of data privacy. Solon Barocas and Helen Nissenbaum identify the risk of a “tyranny of the minority” in big-data analytics when “the volunteered information of the few can unlock the same information about the many [22]. More recently, Salome Viljoen emphasizes the importance of a relational theory of data governance [263]. As Viljoen explains, dataflows entail not only ‘vertical’ relations between a particular individual and a data collector, but also ‘horizontal’ relations between the individual and others sharing relevant population features. Viljoen focuses on the manner in which informational infrastructures rely on group classification to make sense of individuals by taking a ‘relevant shared feature’, generating a prediction based upon that shared feature, and then applying this prediction. When those shared features are derived from synthetic

data, decoupled from any real context and perhaps even specifically created to rectify gaps in representation, we hand power to those creating that synthetic data. We risk imposing designers’ decision-making in lieu of and upon those least likely to have been represented and most likely to be harmed by both the diversity-washing and the side-stepping of consent. After all, if they were represented or able to consent in the first place, there would be no need for additional synthetic data. Data minimization and lineage principles [141] are a first step towards mitigating this issue by requiring documentation and its requisite transparency into where data has come from. The need for this is also readily apparent when considering contexts such as the EU’s Right to be Forgotten [221], where synthetic data further complicates the ability to be removed from a dataset. In making it harder to decouple data from its context through the use of synthetic data, there is an avenue for mitigating consolidation of power and ensuring consent.

Implications for Individual-Control Tools

Many of the risks discussed in this paper propagate from the centralized decision-making nature that synthetic data imposes. Participatory governance structures, as mentioned in Section 4.3, offer a potential solution here — synthetic data could be created to represent concerned groups *by* those self-same groups, re-establishing control and mitigating some concerns around consent and contextualization. Many have called for training data to be restricted to only data that is explicitly consented to be used, though consent is difficult (if not impossible) to establish and propagate over multiple degrees of separation [42]. But synthetic data, when generated to purpose by concerned communities, can provide a potential solution. Models such as those presented in the field of Indigenous Data Sovereignty, where there has been effective push back against external categorization schemas [217] show potential for participatory governance models to address group misrepresentation [10].

Additionally, there are practical considerations that make ‘participatory synthetic data’ an attractive path forward. Both language and computer vision models are beginning to contend with a phenomena commonly referred to as ‘garbage-in garbage-out’ [239, 181]. This refers to the advent of generated data becoming commonplace and public, and the related struggles by those capturing data to differentiate between that data and real data, leading to data generated by a system such as ChatGPT becoming its own training data in the future. Work by Agnew et al. [5] examines the use of these models to replace participants in industry research, highlighting how proposals to do so are motivated by cost reduction and data diversity. They identify these proposals as facing issues in aligning with the values human participants identify as important, specifically including inclusion and representation, necessitating further contextualization and bespoke dataset creation. In such a world, large tech companies may have business motivations for engaging with responsibly created synthetic data, and as demonstrated by Deng et al. [73], methods exist for enabling machine learning practitioners to better contextualize the work they do — a vital aspect of any future responsible synthetic data work.

Further responsible dataset development frameworks that explicitly attend to the partic-

ulars of synthetic data, as well as tooling and practice that examines and makes transparent the provenance of synthetic data, are needed. As examples of this, we propose ‘how could less risky synthetic data be produced?’, as well as ‘how could governance approach consent issues with synthetic data?’ as important future research questions. In future work, we hope to follow the call of Denton et al. [74], contesting machine learning datasets and focusing on the “contingent, historical, and value-laden work practices of actual machine learning researchers” to better understand how the practice of using synthetic data is motivated, the contingent conditions that have lead to its common usage, and the norms and routines that surround it. In so doing, there is the opportunity to survey and better understand the use of synthetic data and create better tools and frameworks for both mitigating its potential for harmful power consolidation, as well as to envision how it can be used as a tool for taking power back [275].

Pulling these strands together, we can now crystallise what the synthetic-data case contributes to the dissertation’s broader argument. The evidence highlights three recurring gaps — what traditional privacy tooling fails to see, why those blind spots arise, and how they propagate into oversight regimes — each of which will resurface in later chapters:

- **Descriptive gap — What we observe.** Synthetic data satisfies per-record privacy tests yet empirically distorts group representation and obscures data lineage, demonstrating that individual-level metrics are an incomplete picture.
- **Explanatory gap — Why it happens.** By design, generators treat each row as fungible; relational patterns that encode social hierarchies are either replicated or amplified, while memorisation artefacts silently re-introduce the very leakage the tool claims to fix.
- **Governance gap — So what.** Enforcement regimes (FTC model deletion, GDPR erasure) and mainstream ML benchmarks presuppose traceable, individually-owned records; once data are rendered synthetic those levers falter, leaving collective harms and accountability unaddressed.

4.8 Conclusion

In this paper, we build on prior responsible dataset development work by focusing on the under-explored impacts of synthetic data on dataset development. Synthetic data will continue to play an ever-increasing role in the training of machine learning systems as real-world data becomes harder to capture, and we must attend to language that paints it as a panacea. We show two examples of the risks of synthetic data, diversity-washing and consent circumvention, and discuss how it is a complicated tool, gravitationally prone to consolidation of power, but with potential for being used to enable participatory governance instead of squashing it. Having shown how synthetic data stretches the individual-control paradigm to

its limits, the next chapter investigates an ostensibly opposite tool — machine unlearning — to test whether revocation mechanisms fare any better when harms are relational.

Chapter 5

Case II: Perceptions of Machine Unlearning

This chapter is based on the paper “User Perceptions of Forgetting in Machine Learning: Exploring Context, Transparency, and Agency”, written in collaboration with Drs. David Piorkowski, Michael Hind, and John Roberts.

5.1 Context

Why Shift from Synthetic Data to Machine Unlearning?

The previous chapter used *synthetic data* to show how a pre-processing solution can satisfy record-level privacy while leaving relational bias and regulatory auditability largely intact. The present chapter moves one step down the ML pipeline — from input substitution to post-hoc erasure. Machine unlearning is marketed as the natural technical realisation of the ‘right to be forgotten’: delete a person’s data, retro-fit the weights, and the system will behave as if the data had never existed [37]. Like synthetic data, unlearning therefore sits at the sweet spot of this dissertation’s analytic lens: it promises to perfect *individual control* through a seemingly straight-forward engineering intervention.

A regulatory magnet.

Supervisory authorities are beginning to signal that honouring an erasure request can require deleting — or retraining — the *models* built on the data, not just the raw rows. The U.S. Federal Trade Commission has already ordered ‘model deletion’ in several biometric cases, including *Everalbum* (2021), *Rite Aid* (2023), and *Ring* (2023), directing the firms to destroy or retrain any models derived from unlawfully collected images or videos [97]. Across the Atlantic, France’s data-protection authority (CNIL) notes in its “Guide d’audit des modèles d’intelligence artificielle” that complying with GDPR Article 17 may oblige

a controller to “re-train or re-initialise” an affected model when simple record deletion is impossible [64]. While these documents stop short of recommending a specific ‘certified unlearning’ algorithm, they illustrate a clear enforcement trend toward *model-level* deletion proof.

A technical reality-check.

Computer-security studies complicate that neat picture, however. Incremental retraining schemes often fail to eliminate first-order gradient influence; shadow and membership-inference attacks can still recover attributes of the ‘forgotten’ [54]. More fundamentally, unlearning targets *parameter attribution*, not *population inference*. Correlations about small social groups, like associations between postcode and default risk, remain baked into the weight matrix even when every row from a dissenting individual is scrubbed [206]. Thus the tool foregrounds the same tension highlighted by RQ1 and RQ2: *does a mechanism optimised for per-row influence also mitigate relational harms?*

An empirical blind spot.

Existing evaluations measure success via statistical metrics — privacy leakage, utility drop, convergence time — but rarely ask the people whose data fuel the models how they understand or value the offer of algorithmic erasure. Do data subjects treat unlearning as genuine empowerment, symbolic reassurance, or a potential collective threat if it degrades fairness for the remaining population? This perceptual layer is crucial: a right that cannot be understood or meaningfully exercised risks becoming another artefact of ‘privacy theatre’ [247].

To surface those situated judgments, Chapter 5 adopts a scenario-based elicitation with heterogeneous technology professionals. The method lets us vary two axes left underspecified in technical papers: **(i)** institutional trust (in this case, comparing food delivery vs. insurance vs. housing) and **(ii)** information state (transparent vs. uncertain vs. breached). By asking participants when they would *opt out*, *delete*, or *unlearn*, we capture the decision thresholds that individual-control remedies must cross to be seen as worthwhile.

Link to research questions.

- **RQ1 — Descriptive/Technical limits.** Our findings reveal the real-world conditions under which users invoke unlearning, setting an empirical lower-bound on any claimed per-row privacy benefit.
- **RQ2 — Relational spill-overs.** Participants’ concern for downstream fairness, even after personal erasure, illuminates how unlearning can satisfy the letter of individual control while leaving group-level harms intact — echoing the representational gaps charted in Chapter 4.

This chapter therefore extends the dissertation’s through-line in three ways:

1. It shows how the locus of individual-control tooling moves from data inflow (synthetic generation) to model weight adjustment (unlearning), yet the relational blind spot persists.
2. It supplies the stakeholder perceptions needed to evaluate governance feasibility — an aspect regulators seldom test when embracing new technical fixes.
3. It prepares the ground for Chapter 6, where we interrogate evaluation benchmarks themselves: if neither pre-processing nor post-hoc erasure fully solves relational harm, perhaps the metrics guiding deployment require contextual overhaul.

In short, unlearning is the next logical probe of the *individual-control* $\leftrightarrow$ *relational-harm* tension, allowing us to triangulate technical limits, user expectations, and policy ambitions within a single empirical frame.

5.2 Introduction

The “right to be forgotten” (RTBF) originated as a reputational remedy but has since become a central plank of data-protection law. It first entered positive law at scale with the European Union’s General Data Protection Regulation (GDPR), in force since May 2018, which entitles any individual to request erasure — and further dissemination bans — whenever data are no longer necessary, consent is withdrawn, or processing is unlawful [221]. The California Consumer Privacy Act (CCPA) followed in 2020 with a narrower, state-level analogue: Californian residents may compel deletion of personal information *held by a covered business*, subject to explicit carve-outs for security, contract, and research uses [266]. Thus, while both regimes aim to restore individual control, the GDPR operates extraterritorially and across all sectors, whereas the CCPA applies only to qualifying firms and contains broader exemptions. Legislatures elsewhere have adopted comparable erasure-style provisions — Argentina’s Personal Data Protection Act, Brazil’s Lei Geral de Proteção de Dados, and Japan’s Act on the Protection of Personal Information among them — each echoing the core GDPR logic while adapting it to local constitutional and commercial contexts [231].

Machine learning complicates the RTBF, as deleting data from a system does not necessarily remove its residual effects on a model [84]. Even after data is erased, models may retain patterns, insights, or behaviours influenced by that data. In other words, deleting a subjects’ data may be insufficient; prior work has show that even if that data is deleted from the system, the model may still reflect their personal information and habits [62]. Thus, it can be difficult for data subjects to understand how their data is being used and the lasting effects of that data on a model’s underlying functionality (including after deletion). This raises critical questions about the feasibility and scope of the RTBF in AI systems [54, 258],

particularly concerning how much control data subjects truly have over data that may leave enduring imprints on models [207].

The development of techniques such as ‘machine unlearning’ [37] hopes to mitigate some of these concerns. Machine unlearning refers to the process of removing or mitigating the impact of data from a machine learning model so that it operates as if the data had never existed [232]. Machine unlearning is still a developing field, and it is not yet clear how effective it will be in practice, or whether it will be possible to revert an AI system to a state that is truly as if the data had never existed [226, 176], but it provides a potential pathway for extending the RTBF into the space of algorithmic systems. There are also legal and policy questions surrounding the RTBF in the context of AI. For example, it is not yet clear what the standard should be for determining when data is no longer needed or is causing harm, and how this should be applied in the context of AI systems [88, 264].

Given the evolving legal and technical landscape, it is important to examine how and under what conditions data subjects wish to exercise the RTBF in the context of machine learning systems. We specifically focus on how data subjects perceive their data being used within in an AI system, and how that affects their preferences around exacting their right to be forgotten. To better understand this, we undertook a scenario-based elicitation activity with 21 participants. Building on prior work in HCI/CSCW using elicitation activities and design scenarios, this method forefronts context, and allows for inductive, interpretivistic, qualitative analysis — a goal of this formative work. As explored in Section 5.3, privacy is a contested, intrinsically contextual concept. Interpretive qualitative research, as Dourish [78] describes, is richly detailed and context-specific, and as Soden et al. [245] articulate, is “valuable for thinking through fundamental concerns that arise when humans and computers interact”, focusing on making specific facets of life legible and understandable. The form of qualitative research we undertook, a scenario-based elicitation activity, is well suited to investigating particularly context dependent scenarios, and allowed for a formative exploration of the particular concerns and interactions that our participants perceived when considering how and when they would want to be forgotten from various AI systems.

In this particular elicitation activity, we asked participants to rank their level of comfort across a set of scenario modifications for three base scenarios and interviewed participants to shed light on our driving research motivation: What considerations do users make when considering having their data be forgotten from a machine learning system? Our interviews focused on how, why, and to what extent study participants wanted to be forgotten across the scenarios. We found participants viewed the RTBF as a mechanism for enacting agency, addressing relational and societal concerns, and re-establishing trust when breached (e.g., through non-consensual data egress). Their preferences were shaped by downstream impacts on others, sensitivity to context, and the use case at hand. We present implications for design from this work, and hope that it sets the stage for future generalizable work through context-friendly large-N participant tools such as factorial vignette surveys [178] to further guide development and regulation around machine unlearning and the .

5.3 Related Work

In this section, we elaborate on how prior and related work in privacy, trust and HCI is foundational to this study, and on how we build upon that. We detail how privacy has been framed as essentially contested and contextual, how user trust in data handling and data use in machine learning systems is similarly predicated on context and verification of expectations, and how HCI has handled the varied nature of privacy preferences and priorities and tackled privacy-sensitive design and user privacy education (from literacy and paradoxes to nudges and nutrition labels). This study builds on these pillars, as we aim to understand how different contexts affect data subjects preferences around having their data be forgotten from machine learning systems, and they guide our choice of engaging in a scenario-based elicitation method. We expand on the prior work by specifically focusing on machine learning systems and, even more narrowly data subjects' preferences around specific methods of effecting control over their privacy in these systems through exercising the .

Privacy — Contested and Contextual

Privacy is an ‘essentially contested’ concept [189]: its meaning shifts across cultures, disciplines, and practical settings. In one context it is framed as an individual’s authority to control data collection and use [99]; in another it is treated as a safeguard for autonomy and dignity, a pre-condition for self-development and free expression. Still other contexts cast privacy primarily as a mechanism for preserving social harmony and regulating interpersonal boundaries [139]. Such plurality is not merely a clash of academic camps; it reflects the situated character of privacy judgments. Shilton and Martin show that people weigh disclosure differently depending on data type, purpose, and setting [179], and later work demonstrates that these preferences evolve with life experience and shifting risk–benefit perceptions [180]. Social norms, regulatory regimes, and the technical affordances of platforms further shape what counts as ‘private’ in practice [121]. Throughout this dissertation we therefore treat privacy not as a single right but as a moving boundary that is negotiated — sometimes contested — within specific sociotechnical arrangements.

In evaluating privacy, it is critical to account for the social and contextual dimensions of information sharing, as well as the norms that govern these interactions. Nissenbaum’s framework of ‘contextual integrity’ [196] provides a foundational lens for understanding privacy as embedded in specific contexts. Contextual integrity highlights four key elements: (1) the context (personal, social, cultural, or organizational settings), (2) the types of information shared, (3) the roles of actors involved in the exchange, and (4) the principles governing data flow. Privacy violations occur when there is a mismatch between the expected norms in a given context and actual practices of data handling.

Our work builds on Nissenbaum’s framework, adding depth to the notion of contextual integrity within machine learning systems. These systems create new contexts where personal data impacts not only individual users but also broader groups and downstream systems. For example, participants in our study reflected on how their decisions to opt out, delete, or

unlearn data might influence others who remain part of the system. This highlights privacy not just as an individual concern but as one that is relational and collective, involving trade-offs between personal comfort, system utility, and societal outcomes.

As prior research has shown, privacy preferences are shaped by a combination of individual needs and collective obligations [180, 139]. Respecting and understanding these preferences is complex, particularly in the context of technology, where data handling practices may conflict with data subjects' expectations [121]. By foregrounding both individual and relational aspects of privacy, our study contributes to a broader understanding of how data subjects navigate the social and systemic dynamics of machine learning systems.

Connection to dissertation: If privacy is negotiated within context and relations, a purely record-level 'forget me' remedy risks missing the social fallout we pursue in later chapters.

Trust — Data Subjects and Machine-Learning Systems

Like privacy, *trust* is conceptually slippery. In sociology and social psychology it names a relational stance — *willingness to be vulnerable* to another party under conditions of risk. Jen King's dissertation, for example, treats trust as something that exists *between* individuals and the institutions that hold their data, contingent on power asymmetries and the possibility of harm. Computer-science discourse often shifts the focus from this relational notion to a system property called *trustworthiness*: whether a technical artefact meets stakeholders' expectations in a verifiable way. The EU's 'trustworthy AI' agenda crystallises that move, listing four normative pillars — respect for human agency, prevention of harm, fairness, and explainability — and Kaur et al. add a fifth, *acceptance*, to underline the socio-technical dimension of adoption [154].

Across both framings, certain design requirements recur: human override or recourse [71], robust data-governance processes [118], transparency about how decisions are reached [77], non-discrimination [103], promotion of social well-being [244], and clear accountability trails [276]. Empirical work suggests that people *say* they would trust an AI system more when such conditions are in place, but the underlying studies often measure stated comfort rather than behavioural commitment and rarely engage deeply with trust theory [238]. The gap matters: users routinely report that transparency is important yet decline to read or act on lengthy disclosures when they are offered — a pattern documented in a UK Cabinet Office study on public-sector algorithms [87].

Two structural factors shape trust decisions in data practice. First is **institutional trust**: which kinds of organisations are deemed competent and benevolent custodians of personal data. Survey work finds that financial institutions and hospitals score highest, government next, and social-media platforms lowest [261, 178]. Second is **data modality**: users differentiate sharply among data types, rating health and financial records as most sensitive, social-graph data next, and entertainment preferences least so [16, 68]. Our scenario design leverages both dimensions by crossing three domains (food delivery, insurance, landlord decision-making) with varying levels of disclosure and control.

Consistent with prior work, we find that perceived boundary violations — especially data egress to unknown third parties — trigger the most aggressive responses, including demands for deletion or machine unlearning. These reactions reinforce the idea that trust in AI systems is not a stable attitude but a situational calculus shaped by domain stakes, perceived agency, and institutional reputation — variables we explicitly manipulate in the study and return to in the design and policy implications.

Connection to dissertation: Trust hinges on transparency and control; unlearning promises both, letting us test whether technical levers actually deliver the relational trust participants demand.

Users — Privacy Literacy, Paradoxes, and Design Interventions

Human-computer-interaction research treats end-user privacy as both *experience* and *governance*. Early work framed privacy as a design resource: interfaces that respect disclosure boundaries and surface data flows improve usability and foster trust [142, 151]. Subsequent studies argued that such respect hinges on two levers — **control** (letting people veto or curate data capture) and **transparency** (showing what is captured, when, and why) [265, 230]. Yet the empirical record is messy. Some users will trade privacy for convenience, others will not; preferences even for the same individual vary by task, context, or life stage [190]. The result is the well-known ‘privacy paradox’, in which stated concern and observable behaviour diverge.

Jen King’s work clarifies why such divergence is not mere hypocrisy but a *structural* outcome. Her dissertation shows that lay people use *institutional trust* as a cognitive shortcut: they will share data with an organisation they perceive as competent and accountable even when they cannot parse the technical details [157]. King’s experiments with ‘privacy nutrition labels’ likewise demonstrate that simplified disclosures help only when users believe the label issuer is worthy of trust and when the label appears at the moment of choice, not buried in a settings menu. Put differently, literacy tools and nudges are necessary but insufficient; they must be coupled with signals that the collecting institution merits reliance under risk.

The broader HCI literature echoes that diagnosis. Users often lack a mental model of how data are processed [256]; increasing transparency can even backfire if the information is illegible or overwhelming [2, 119]. Attempts to close the gap — nutritional labels, permissions nudges, just-in-time prompts — move knowledge closer to the point of action, but recent critiques note that ‘privacy literacy’ by itself does not automatically yield agency. Trust, institutional reputation, and perceived vulnerability still shape decisions, mirroring patterns observed in AI fairness and trustworthy-AI research. The scenario design in Chapter 5 builds on King’s insight: by varying both the level of information and the identity of the data-holding institution, the study can observe whether users default to literacy cues, relational trust, or both when deciding to opt out, delete, or invoke machine unlearning.

Connection to dissertation: The literacy-action gap mirrors the consent-control gap: unlearning offers a high-salience act whose perceived power we empirically probe.

Contexts — Scenarios, Elicitations and Situational Values

Pommeranz et al. [210] detail how contested values such as privacy are situational values, and when seeking to understanding situational values, “standard knowledge gathering methods (e.g. interviews or surveys) lack in giving the real life context needed for people to understand and express their values”. This understanding is furthered by prior privacy preference work that utilized factorial vignettes and highlighted the “relative importance of particular contextual factors and information uses” [178]. This also follows calls by the likes of Barkhuus to question the “viability of obtaining universal answers in terms of people’s ‘general’ privacy practices”. [20]. Context-awareness has been much studied in fields such as ubiquitous computing, with scenario-based approaches frequently preferred when changes in context are expected or anticipated [235, 236, 241]. In HCI, design workbooks and cards are frequently used as elicitation studies, including Wong et al. [277] and Friedman & Hendry [106], affording rich participant responses, alongside uses of other devices such as a probes and participatory engagements which seek to ground designs in the contexts they will be placed within and affect [282, 53, 159]. We build on Naeini et al. [86], who iteratively elicited factors to compose privacy and security ‘nutrition labels’ for IoT devices, as well as Yao et al. [286]’s analysis of privacy concerns from the perspective of different users in a role play activity, and Psychoula [212]’s work on privacy modelling for assisted living within smart homes, which used a similar scenario-based elicitation method. This paper uses scenarios in an elicitation activity to elicit users perceptions and motivations across different scenario contexts.

Connection to dissertation: Contextual elicitation uncovers the situated trade-offs that aggregate tools (synthetic data, unlearning, context metrics) routinely gloss over.

5.4 Method

We conducted a scenario-based elicitation activity, interviewing 21 participants from an international technology company. While broader recruitment could have offered additional perspectives, this approach was well-suited to the study’s formative goals at making legible the type of considerations data subjects face and the need for depth in responses within the project’s constraints. This participant pool was selected for its diverse roles and expertise levels, which allowed for nuanced perspectives on AI systems without extensive user education. Table 5.1 provides aggregated demographic and technical expertise data from a screening questionnaire. Care was taken to recruit across age, gender, and levels of AI knowledge, though most participants likely had above-average familiarity with AI due to their professional context. This higher baseline of knowledge supported the study’s goal of exploring complex privacy considerations within AI systems.

Participants provided written informed consent and were interviewed for 60 minutes, with compensation equivalent to \$25 via an internal company rewards system. The semi-structured interviews focused on understanding participants’ rankings of scenario modifica-

Table 5.1: Participant Demographic Data

	Number of Participants	Percent
Gender		
<i>Male</i>	9	43%
<i>Female</i>	11	52%
<i>Other</i>	1	5%
Age		
<i>18-25</i>	3	14%
<i>26-35</i>	8	38%
<i>36-45</i>	5	24%
<i>46+</i>	5	24%
AI understanding		
<i>Passing</i>	4	19%
<i>Conceptual</i>	12	57%
<i>Deep</i>	5	24%
Currently employed in AI		
<i>Yes</i>	7	33%
<i>No</i>	14	67%
Job function		
<i>Marketing</i>	6	29%
<i>Design</i>	2	10%
<i>Software Engineering</i>	4	19%
<i>Research</i>	4	19%
<i>Operations</i>	2	10%
<i>Sales</i>	3	14%

tions and their decision-making processes. Interviews were conducted via Webex, with screen sharing to present scenarios and modifications. Interviews were transcribed using Webex's transcription feature and coded inductively by two researchers. Using a consensus coding approach [182], themes were identified to capture commonalities and differences in participants' decision-making processes. These themes formed the basis for the findings presented in the Results section.

The elicitation consisted of participants engaging with three scenarios representing varying levels of institutional trust and data sensitivity, derived from the AI Incident Database [183]:

1. **Food Delivery:** An app uses a machine learning system to suggest orders and targets ads based on phone data.
2. **Insurance:** A system determines rates using credit history, income, and public online data to generate risk scores.
3. **Landlord:** Smart home data and online information are used by a system to calculate annual rent updates.

Participants were also presented with scenario modifications detailing different knowledge states (e.g., “You don’t know what data the system was trained on”). These modifications, derived from explainable AI and privacy literature [169], were categorized into five key questions about data use. Each question included three variations: unknown, positive, and negative. Participants ranked these modifications from most to least comfortable, which served as a foundation for the decision-making activity.

Participants were then introduced to three actions they could take regarding their data:

- **Opt out:** Stop using the system; no further data collection occurs.
- **Delete:** Previously collected data is removed; the participant is opted out.
- **Unlearn:** The system is updated as if the participant’s data had never been used.

Participants were given a brief description of each of these three actions, and unlearning was presented as a realistic option to allow participants to consider when they would want to use it. Participants indicated at which point on their ranked list of scenario modifications they would choose to take each action for each scenario, drawing lines to represent these decisions. For example, a participant might draw the ‘opt-out’ line under “You don’t know how accurate system outputs are or how they’re measured” and the ‘delete’ line under “You don’t know the exact logic used by the system”. These decisions were discussed in detail to uncover motivations and comparisons across scenarios. See Section 5.1 for visual representation of ranking and decision-making.

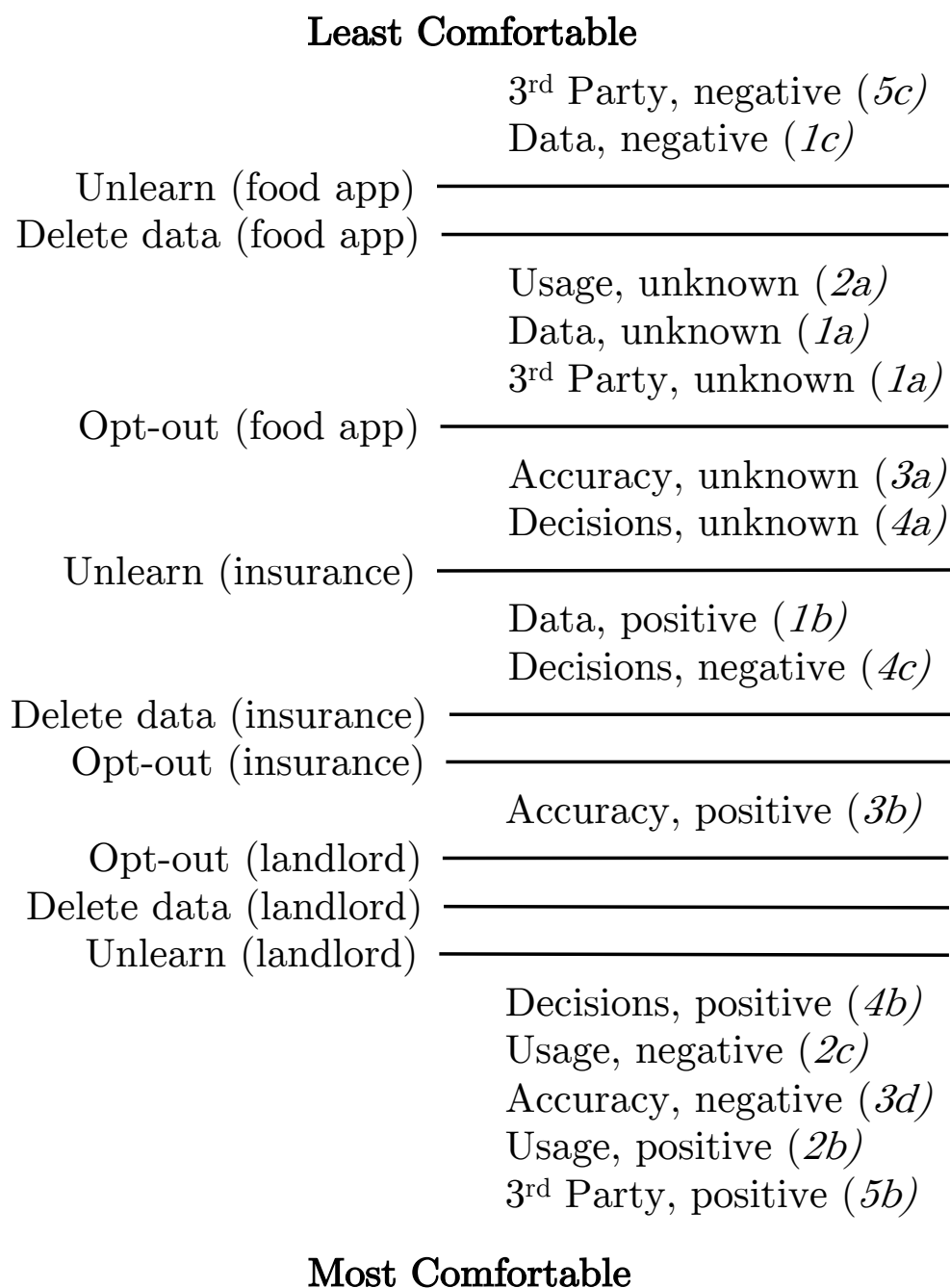


Figure 5.1: Decision-Making Reference Figure, Participant U

Participants ranked information stages and then drew lines on those ranked lists to represent three different decisions (opt out, delete, unlearn) they could take across three different scenarios of data usage in an AI application (landlord, insurance, food app).

1. What kind of data was the system trained on? (*Data*)

- a) You don't know what data the system was trained on.
- b) You know the system is trained on a narrowly scoped set of data (only data they can demonstrate is needed).
- c) You know the system is trained on a broadly scoped set of data (anything they can access or purchase).

2. How are system outputs used? (*Usage*)

- a) You don't know how the systems outputs are used.
- b) You know that the system outputs a set of factors to be considered by a human agent to make decisions.
- c) You know that the system outputs a decision automatically.

3. How accurate/precise/reliable are the results? (*Accuracy*)

- a) You don't know how accurate system outputs are or how they're measured.
- b) You know that the system is as accurate as the previous non-machine learning system.
- c) You know that the system's accuracy is measured against goals set by the designers.

4. Why is this the result? (*Decisions*)

- a) You don't know how the system makes decisions from input data.
- b) You know the exact logic used by a system to make decisions
- c) You know the system is a black box, decisions are not interpretable.

5. How is the system/the data secured? (*3rd Party*)

- a) You don't know what happens with the system/the data outside of the intended usage.
 - b) You know that the company doesn't sell to third parties and has a stellar reputation for data security.
 - c) You know that the company sells data/system results to third parties and has a history of data breach.
-

Figure 5.2: Base Questions and Resulting scenario modifications

Participants ranked these information stages and then drew lines on those ranked lists to represent three different decisions (opt out, delete, unlearn) they could take across three different scenarios of data usage in an AI application (landlord, insurance, food app).

5.5 Results

Overview

This section presents findings from participant interviews, organized around key themes identified during analysis: perceptions of control, contextual sensitivities, the impacts of their decisions, and trust in AI systems. Participants' responses revealed nuanced decision-making processes regarding when and how they felt comfortable with their data being forgotten in AI systems.

Perceptions of Control

Participants' perceptions of control played a significant role in shaping their comfort with AI systems and their data retention preferences. Across the interviews, the themes of agency, transparency, and explainability emerged as critical factors in participants' assessments of the systems they encountered. This section examines how these factors influenced participant comfort and their sense of control over AI systems, with implications for system design.

Agency

Participants consistently emphasized the importance of agency — the ability to directly influence system behaviour — as a primary factor in their comfort with AI systems. Scenarios that restricted participants' control over their personal data were viewed negatively, while those that provided mechanisms for direct interaction or influence fostered greater comfort and trust.

For example, Participant L expressed discomfort with scenarios where they felt unable to influence the system, stating, *"I don't know what data the system was trained on. That's really bad — again, I, at that point, to whatever extent I can choose what I do with that information, I've lost any sort of agency."* Similarly, Participant I drew a connection between agency and human involvement, explaining, *"I wouldn't really care about how the data was used or how bad it is, because you can also negotiate with the person and say, look at this. This is wrong."* Such responses underscore the importance participants placed on having tools or pathways to shape system behaviour.

Explainability was also closely tied to agency, as participants described it as a critical factor enabling them to assess and act on system decisions. Participant A highlighted how system-level explanations could foster trust, stating, *"The company's making an intentional decision to explain and make it understandable to me, I like that and trust them more."* They further noted that explainability enabled intervention when needed, elaborating, *"if they could say the reason this happened is because of this... then maybe they could change it if it went wrong."*

Transparency

Transparency emerged as a critical enabler of participants' sense of agency, providing the information needed to make informed decisions and assess system behaviour. While agency emphasizes participants' direct influence, transparency centres on their ability to understand how and why decisions are made, which is a prerequisite for exercising meaningful control. For instance, Participant U emphasized the importance of understanding how their data was used, particularly in high-stakes contexts, stating, *"for example, while renting or getting credit, I should be aware that my income contributes to these methods. I think I should be aware of the main function of use and so on. So, if I don't know, at that point, I'm uncomfortable."* Similarly, Participant P expressed the necessity of transparency in sensitive domains, explaining, *"with the insurance, I'd want it to be a little more transparent on what's being weighted so I know whether to be comfortable."*

Transparency was seen as a form of indirect control, compensating for the absence of direct agency. Participant G framed transparency as a trade-off, stating, *"when I see that's not possible, I am kind of saying that okay, I cannot have that, but I would be happy to compromise and instead know that they have enough information about my demographics, where I come from, what's my education and stuff like that."* By providing clarity, transparency allowed participants to navigate and evaluate systems more effectively, even when direct intervention was not feasible.

Conversely, a lack of transparency often led to discomfort and mistrust. Participant E articulated this, stating, *"If I'm using a system, almost no matter what the system is, if you then find out that it's unclear what the input data is doing. That's kind of like the yeah, that's too far."* This sentiment highlights the critical role of transparency in mitigating discomfort and fostering user confidence.

Contextual Sensitivities

Participants' decisions around opting out, deletion, and unlearning were strongly shaped by contextual factors. These sensitivities manifested in two primary ways: (1) the domain of data usage, which influenced participants' perception of risk and relevance, and (2) their reasoning strategies for navigating perceived gaps in understanding or control over system behaviour. The first theme, domain sensitivity, highlights the heightened caution participants exhibited in scenarios involving critical areas such as housing or insurance, compared to lower-stakes contexts like food delivery apps. The second theme, gap-motivated reasoning, underscores how participants reconciled their comfort (or discomfort) with AI systems based on perceived limitations in transparency or alignment with personal values. Together, these findings reveal how the interplay of domain-specific considerations and reasoning strategies shapes users' actions and attitudes toward AI systems.

Domain Sensitivity

Participants' perceptions of data use varied significantly across different domains, reflecting how context shaped their comfort and decision-making processes. Sensitive domains, such as insurance and housing, elicited heightened concern, while lower-stakes contexts, such as food delivery, often led to apathy or reduced scrutiny.

In high-stakes contexts, concerns about data misuse often escalated into acute discomfort. Participant A described how trust violations, such as third-party data breaches, could trigger a desire for complete erasure: *"if I said delete my data, but then later I learned that my insurance company, my landlord, or the AI system that they use has been selling my data to 3rd parties and has a history of a data breach I would be like, I'm gone ... I want everything to be like I never existed."* This response illustrates how domain-specific sensitivities can exacerbate mistrust, particularly in scenarios where participants feel their data is being mishandled.

Domains perceived as lower stakes, such as food delivery, frequently inspired apathy toward data use. Participant F explained, *"with the food app, like I said, I really don't care that much about it, about how they're using the data. I could see how it can still be abused but realistically a lot of the items below that line aren't of as much concern with that data but with the other two I feel like they're just such egregious abuses of the data."* However, even within these less sensitive domains, participants occasionally raised concerns about potential abuse or secondary impacts.

The variation in sensitivity across domains also influenced participants' choice of action when exercising their right to be forgotten. Participant K, for example, noted a lack of concern regarding data deletion in the food delivery scenario, explaining, *"I don't see a relative impact between opt-out or delete or unlearning for the food scenario considering the fact that I don't really care how they measure."* This highlights the interplay between domain sensitivity and the perceived necessity of different interventions, with participants calibrating their actions based on the stakes involved.

Gap-Motivated Reasoning

Participants frequently described their comfort with AI systems as being shaped by perceived gaps in their understanding of either the system's decision-making process or its outcomes. These gaps were framed as either tolerable or intolerable, depending on the participant's expectations and self-perceived ability to evaluate the system. For example, Participant C highlighted their indifference toward understanding a system's internal logic, stating, *"the exact logic used by a system to make decisions, I don't know why I should be aware of that. As long as the system works, if I don't know the exact logic, well, why would I know? I'm a normal data subject."* Similarly, Participant U prioritized outcomes over process transparency, explaining that they were *"less concerned about the black box model and more about how the people use the outputs of the model."*

However, other participants expressed significant discomfort when gaps in their under-

standing aligned with contexts they deemed high-risk. Participant T described a specific concern about third-party access to data, saying, *“I worry that if I don’t know that [how the data is used] that they could go to a separate merchant, which has a history of data breaches themselves.”* This anxiety was particularly pronounced in scenarios where participants lacked the expertise to assess the system’s reliability. As Participant F explained, *“with the outputs, I think for me at that point it’s beyond my level of expertise, I don’t know that I would be qualified to make a decision.”*

Impacts of Decisions

Participants frequently contextualized their decisions around opting out, deletion, and unlearning by reflecting on the broader implications of these actions. Two primary themes emerged: (1) the perceived downstream impacts of data removal on fairness, utility, and societal outcomes, and (2) how system-level data usage influenced participants’ trust in the system and willingness to remain engaged.

Downstream Impacts

In considering downstream impacts, participants expressed concerns about the potential for their decisions to unintentionally harm others or disrupt societal systems. These reflections often centred on sensitive domains like insurance and housing, where participants weighed the ethical trade-offs of unlearning.

Participants frequently considered the broader consequences of their decisions to unlearn or delete their data, particularly in sensitive domains like housing and insurance. These downstream impacts were often framed around fairness, societal utility, and ethical trade-offs. For example, Participant F described choosing to unlearn their data because *“they could be causing harm to others, and probably in a much larger scale just because insurance companies tend to cover many more people than like a single landlord would cover.”* Participant Q echoed this sentiment, explaining that unlearning was necessary *“because it could impact other people’s lives and if it’s maybe unfair.”*

Some participants reflected on the positive societal implications of their data being used responsibly. Participant H justified allowing their data to remain, reasoning that it might support *“measuring things like security or energy output, things that we would be for the greater good of the apartment complex.”* Similarly, Participant L pointed to specific scenarios where highly personal data might still be used to benefit others, emphasizing the importance of an *“opt-in to [unlearning] instead of a blanket default.”*

System Data Usage

Regarding system data usage, participants highlighted the perceived stakes associated with different domains, demonstrating a nuanced understanding of how data removal might affect both system performance and individual outcomes. This theme was particularly evi-

dent in discussions of trust and accountability, with participants linking their comfort levels to the transparency and reliability of system behavior.

Participants demonstrated a nuanced understanding of how their data removal decisions could affect system performance and societal outcomes, with their comfort levels often tied to the perceived importance of the domain in question. For less critical domains, participants expressed a general apathy toward data usage. For example, Participant F explained, *“With the food app, like I said, I really don’t care that much about it, about how they’re using the data. I could see how it can still be abused, but realistically a lot of the items below that line aren’t of as much concern.”*

However, this apathy was contrasted by heightened caution in sensitive domains like insurance or housing, where participants weighed the risks of data misuse more heavily. Participant B compared the stakes between domains, stating, *“I feel like a real estate agent isn’t going to delve into my personal data, and I could move around easier than finding new insurance.”* Similarly, Participant A emphasized the importance of trust in these high-stakes domains, saying, *“If I said delete my data, but then later I learned that my insurance company, my landlord, or the AI system that they use has been selling my data to 3rd parties and has a history of a data breach I would be like, I’m gone ... I want everything to be like I never existed.”*

These decisions were also influenced by participants’ perceptions of how data was being used at a system level. Some, like Participant K, dismissed concerns in lower-stakes scenarios, stating, *“I don’t see a relative impact between opt-out or delete or unlearning for the food scenario considering the fact that I don’t really care how they measure.”* Others, such as Participant G, reflected on the redundancy or accessibility of alternatives in specific domains, saying, *“I’m assuming there are a lot of insurance companies, so even if I get rejected by most of them, I will just go for another, and it’s not like my life depends on it or something.”*

Trust in AI Systems

Trust emerged as a key factor in participants’ perceptions of their data being used in AI systems, influencing both their comfort with data use and their decisions around unlearning, deleting, or opting out. Participants frequently associated trust with transparency and the ability to understand system behaviour. Participant A emphasized how intentional efforts by a company to explain its practices increased trust, stating, *“The company’s making an intentional decision to explain and make it understandable to me, I like that and trust them more.”* Similarly, Participant P highlighted the importance of transparency in sensitive domains, explaining, *“With the insurance, I’d want it to be a little more transparent on what’s being weighted so I know whether to be comfortable.”*

Conversely, a lack of trust often led participants to choose more aggressive actions like unlearning. Participant J expressed frustration with undisclosed data practices, saying, *“If they didn’t communicate it to me [selling data], I totally don’t want them to use my data to train, I don’t want them to have anything from me.”* This sentiment was echoed by Participant N, who described a loss of trust as a trigger for comprehensive actions:

“Realistically, from my personal perspective, if I’ve gotten to the point where I don’t trust the company and they give me all the levers to forget myself, I’ll probably just go ahead and do that.”

5.6 Discussion

This study examined how participants perceive and navigate the *right to be forgotten* in AI systems. Through a scenario-based elicitation activity, we explored the contextual factors that influence their decisions about when and how they would like to be forgotten, particularly focusing on machine unlearning as a means of achieving this. Four key themes emerged: perceptions of control, contextual sensitivities, impacts of decisions, and trust in AI systems. This discussion connects the findings to existing literature and considers their implications.

Perceptions of Control

Participants’ perceptions of control emerged as a critical determinant of their comfort with AI systems and their decisions to engage with or withdraw from such systems. This theme encompassed two key dimensions: agency — the ability to influence system behaviour — and transparency, which participants viewed as both an enabler of agency and a stand-alone factor in shaping trust and comfort.

Agency and System Design

Agency was central to participants’ assessments of AI systems, particularly in scenarios that restricted their ability to influence outcomes. Participants consistently described discomfort with systems that limited their control over personal data or decisions. This aligns with prior work framing privacy as a means of asserting control over one’s information [151, 230]. Participants’ preferences for agency often translated into a desire for mechanisms that allowed them to interact with or override system decisions. For instance, the inclusion of a human-in-the-loop was frequently cited as a means of enhancing agency. While this preference aligns with prior work highlighting the value of human oversight in AI systems [234], it also underscores a tension between participants’ comfort with automation and their need for influence. Participants who preferred human involvement viewed it as a safeguard against errors or biases in automated decision-making, reflecting concerns about the opacity of machine learning systems. However, this preference does not necessarily align with the practicalities of human-in-the-loop systems, which often rely on human oversight rather than direct engagement with users. Future design efforts should distinguish between these roles and consider the conditions under which human oversight adds value without introducing new biases or inefficiencies.

Transparency as an Enabler of Agency

Transparency was repeatedly described as a critical enabler of agency, allowing participants to make informed decisions about their data and its usage. Participants associated transparency with trust, emphasizing the importance of clear communication about system behavior, data usage, and decision-making processes. These findings echo prior research identifying transparency as a cornerstone of trust in AI systems [105]. Participants viewed transparency as particularly important in high-stakes contexts, such as housing and insurance, where the perceived risks of data misuse were greater. For example, participants emphasized the need to understand what data was being collected, how it was being used, and how it influenced system outputs. Without this information, participants frequently expressed discomfort and a lack of trust in the system. This aligns with prior work highlighting the role of explainability in fostering user trust and engagement [12]. However, transparency alone was not always sufficient to alleviate discomfort. Some participants described feeling overwhelmed by the complexity of system processes or unsure of how to interpret available information. This suggests that transparency efforts must be coupled with clear, accessible explanations that enable users to act on the information provided. Design approaches such as simplified data visualizations or tailored explanations could address this challenge, empowering users to navigate complex systems more effectively.

Enabling Trust in Design

Participants noted that when transparency revealed flaws or inconsistencies in system behaviour, trust often diminished significantly, prompting them to favour more decisive actions, such as unlearning. This highlights the need for transparency to go beyond merely disclosing information; it must foster actionable understanding. Providing users with real-time feedback about the effects of their actions or offering interactive tools to explore system processes can bridge this gap, empowering users to make informed decisions. These interventions are particularly important in scenarios where trust and agency are deeply intertwined, as they ensure users feel both informed and capable of influencing outcomes.

Transparency, while foundational for enabling agency, also stood out as a stand-alone factor shaping participants' perceptions of fairness and reliability. When trust in the system was compromised — whether due to perceived opacity or mismanagement — participants were more likely to advocate for unlearning as a way to reclaim their agency. Unlearning was viewed as a proactive means of asserting control over their data and addressing a breach in trust, extending beyond the limitations of simple data deletion. This response sheds light on when and why users might escalate their privacy actions, particularly in high-stakes contexts where trust is critical.

For system designers, these findings emphasize the importance of integrating transparency and agency-enhancing features into AI systems. Additionally, by providing the option for forgetting-via-unlearning as a mechanism for regaining trust, designers can signal a commitment to user autonomy. Even if the goal is to minimize the need for unlearning,

its availability may foster trust and empower users, creating systems that prioritize fairness, accountability, and user confidence.

Contextual Factors in Data Retention Decisions

Participants' decisions regarding data retention, deletion, and unlearning were heavily influenced by contextual factors. These factors included the sensitivity of the domain in which their data was used and the reasoning strategies they employed to navigate gaps in their understanding or control over AI systems. This section explores how domain-specific concerns and gap-motivated reasoning shaped participants' actions, and implications for system design.

Domain Sensitivity and Perceived Stakes

Participants consistently expressed that the perceived stakes of a domain influenced their comfort with data usage and their willingness to engage in forgetting actions. Sensitive domains, such as housing and insurance, elicited heightened concern, while lower-stakes contexts, like food delivery apps, were often met with apathy. This aligns with prior work demonstrating the role of contextual integrity in shaping privacy preferences [196].

For high-stakes domains, participants described heightened anxiety about potential data misuse and its impact on fairness and access. For example, participants were particularly concerned about third-party data sharing or breaches in sensitive contexts, echoing prior research linking privacy concerns to perceptions of vulnerability and harm [163]. This sensitivity often led participants to advocate for stronger protections, such as unlearning, in scenarios where they felt their data could be misused.

In contrast, participants exhibited apathy or reduced scrutiny in low-stakes contexts. For instance, data usage in food delivery apps was frequently perceived as inconsequential, with participants expressing less concern about how their data was used or stored. However, even in these cases, participants frequently turned to unlearning in scenarios where there was major data breach.

A context-sensitive approach to unlearning could provide opportunities for system design. In assuaging anxiety, robust transparency mechanisms that provide clear explanations of data usage and safeguards against misuse could help prevent users from wanting to turn to more computationally intensive methods of forgetting such as unlearning. In situations where participants cared less about how the data was used, as long as it wasn't abused in breach of the context it was provided or collected for, enabling unlearning is potentially overkill, with it instead being more of a proverbial stick to be utilized against bad actors.

Gap-Motivated Reasoning as a User Strategy

Participants frequently navigated gaps in their understanding or control over AI systems by employing reasoning strategies that reconciled their comfort with the limitations of these

systems. These strategies often involved prioritizing either process transparency or outcome reliability, reflecting participants' differing values and expectations.

For some participants, the perceived opacity of system processes led to a focus on outcomes as the primary metric for evaluating comfort. These participants described a sense of resignation to the inevitability of black-box decision-making, prioritizing results over an understanding of how those results were achieved. This perspective aligns with prior work on the trade-offs between transparency and convenience in user decision-making [190]. Conversely, other participants emphasized the importance of understanding the processes behind system behaviour, particularly in high-stakes contexts. These participants expressed discomfort with scenarios where they lacked the expertise or information to evaluate system reliability, reflecting concerns about accountability and trust. This echoes findings from prior studies that highlight the role of explainability in mitigating user anxiety [105].

To support users navigating these gaps, systems must provide clear, accessible information about their decision-making processes and outcomes. Participants frequently emphasized the importance of actionable transparency — providing not only insights into system behaviour but also tools for acting on those insights. For instance, interactive explanations or real-time feedback mechanisms could help bridge gaps in understanding, enabling users to feel more confident and empowered in their interactions with AI systems. Additionally, systems should consider integrating educational components that address common knowledge gaps among users. For example, offering concise tutorials or contextual help could help demystify complex processes and foster greater trust and engagement. As mentioned previously, these interventions are particularly important in high-stakes domains, where users are more likely to scrutinize system behaviour and demand accountability.

Impacts of Forgetting on Systems and Stakeholders

Participants' decisions to opt-out, delete, or unlearn their data were also shaped by considerations of downstream impacts on system performance, fairness, and broader societal outcomes. These decisions reveal the complexity of balancing individual preferences with system-level implications, particularly in sensitive domains.

Participants frequently reflected on the broader implications of their decisions to be forgotten, particularly how these actions might affect other users or the overall functioning of the system. In sensitive domains, such as insurance and housing, participants expressed a heightened awareness of fairness and systemic impact, aligning with prior work on the social dimensions of privacy [139].

For instance, participants described unlearning as a means of addressing perceived harms caused by biased or unfair system behaviour. This perspective was particularly evident when participants considered how their data might contribute to broader societal inequities. However, participants also acknowledged potential trade-offs, with some expressing concern that unlearning could disrupt system performance or utility. For example, Participant L noted the importance of balancing personal privacy with the potential benefits of data-driven insights.

These findings highlight the need for nuanced system designs that address fairness and utility concerns while respecting user preferences. Participants displayed nuanced and varied understandings of how unlearning could enable greater agency over the systems using their data, while also seeing it as a path towards mitigating harm caused by data misuse.

Implications for Individual-Control Tools

Across participants' narratives, machine unlearning reproduces the three-fold tension already surfaced in Chapter 4. The findings sharpen the dissertation's analytic lens as follows:

- **Descriptive gap — What we observe.** Participants reach for unlearning to eliminate *direct, record-level exposure* — e.g. the fear that a single health record might leak — yet they still worry about higher insurance premiums or rent hikes that arise from biased group statistics. Row deletion therefore leaves supra-individual harms untouched.
- **Explanatory gap — Why it happens.** Unlearning algorithms scrub a data point's explicit influence on model parameters, but they do not remove the population-level correlations from which those parameters draw meaning. As long as the correlation between postcode and default risk persists, the social meaning of one's data lives on.
- **Governance gap — So what.** Erasure rights in GDPR and allied laws align neatly with unlearning APIs, yet regulators possess no metric — or audit log — to verify the tool's collective impact. The oversight blind spot documented for synthetic data thus reappears: compliance can be demonstrated for the individual while group harm goes unmeasured.

Limitations and Future Work

This study offers valuable insights into data subjects' perceptions of the right to be forgotten in AI systems but has several limitations that point to future research opportunities.

First, our participant sample consisted of employees from a single large technology company, likely with above-average familiarity with AI and privacy concerns. While this provided diverse perspectives within a technically literate population, it limits generalizability to broader demographics. Future research should expand to include diverse professions, cultural backgrounds, and levels of technical expertise, particularly those who may face systemic discrimination or harm.

Second, the study framed machine unlearning as a fully viable option, equivalent to opting out or data deletion. While useful for eliciting participant perspectives, this simplified the technical realities, as unlearning is not universally feasible across machine learning architectures [206]. Future work should examine how perceptions of unlearning evolve as the technology matures and its practical limitations become more evident.

Third, our scenarios, while rich in eliciting qualitative insights, were constrained to pre-defined contexts and did not account for broader social or community dynamics. Exploring these dynamics and how they intersect with individual decision-making, especially in high-stakes domains, remains an important avenue for future work.

Finally, while our findings highlight variation in participants' comfort and decision-making across domains, the small sample size limits the extent to which these results can be generalized. Future studies could validate and extend these findings using large-scale surveys or factorial experiments, particularly to investigate how factors such as domain sensitivity, transparency, and perceived agency interact to shape preferences.

Despite these limitations, this study lays an important foundation for understanding the complex interplay of agency, trust, and contextual factors in data subjects' decisions to exercise the right to be forgotten. Addressing these limitations will enable the development of equitable, user-centered AI systems.

5.7 Conclusion

This study provides insight into data subjects' preferences and concerns regarding being forgotten from machine learning systems, as the nascent field of machine unlearning aims to undo the impact of specific data points on a model. We explored how data subjects make decisions about what, when, and why they would like to be forgotten by machine learning systems.

Using a scenario-based elicitation activity, we found that participants:

- Not only consider a system's impact on themselves but also evaluate how their privacy decisions might affect others.
- View agency over the system as a key factor in their comfort with it, often characterizing the as a means of influencing system behaviour.
- Demonstrate sensitivity to the use case and context, particularly around high-stakes domains, while expressing general discomfort with data egress beyond its intended purpose.

Taken together, these insights expose exactly the descriptive, explanatory, and governance gaps outlined in the preceding discussion. They show that individual requests for erasure rarely end at the individual: each deletion decision reverberates across groups and system performance, and the current compliance toolbox offers no fully auditable guarantee that those reverberations are benign. Chapter 6 picks up this thread by examining context-sensitive evaluation metrics for ML systems, asking whether measurement practices can be redesigned to surface the very relational impacts that unlearning-based privacy tools leave hidden.

Chapter 6

Case III: Contextual Evaluations

This chapter is based on the forthcoming paper “Evaluating Large Language Model Adherence to Instructions, Norms, and Rules in Context: A Literature Review”.

6.1 Context

From data *inputs* to model *outputs*.

Chapter 4 and Chapter 5 interrogated the two ends of the individual-control pipeline. Chapter 4 asked whether *better data*, via synthetic generation, could eliminate consent and representation problems at the point of *collection*. Chapter 5 flipped the temporal direction, probing whether *more flexible models*, via machine unlearning, could restore autonomy *after* data have already been ingested. Both studies showed that notice-and-consent tools fracture once the harm surface widens from single records to relational, group, and systemic effects.

Why evaluation becomes the bottleneck.

If neither obtaining ‘perfect’ inputs nor scrubbing undesired memories guarantees downstream safety, then governance must increasingly depend on stringent *evaluation* of model behaviour *in situ*. Large-language models (LLMs) now draft policy, propose medical regimens, and write production code. In each case stakeholders care less about where a particular datum originated and more about whether the *output* (i) obeys explicit instructions, (ii) respects contextual or cultural norms, and (iii) conforms to formal rules (legal, logical, organisational). Failures on any layer can trigger the representational, allocative, and systemic harms mapped in Chapter 2.

The missing middle: context as algorithmic substrate.

Current adherence benchmarks often test these layers in isolation or strip away the situational cues that give norms their force. A system can ace a keyword checklist yet violate the

etiquette of a grief-support forum; it can quote a privacy policy verbatim yet divulge patient data when the same request is paraphrased. Context is thus not cosmetic — it is the substrate that determines which informational flows are appropriate, echoing Nissenbaum’s *transmission principles* in contextual integrity. Without evaluations that foreground this substrate, we risk what Shvartzshnaider and Duddu call ‘contextual-integrity washing’, where systems claim compliance precisely because the tests themselves are de-contextualised [240].

How this chapter closes the loop.

Chapter 6 catalogues — and critiques — the state of LLM adherence benchmarks from three angles:

1. **Explicit, verifiable instructions** (e.g. format, length, keyword, style).
2. **Contextual social and cultural norms** (implicit expectations embedded in stories, dialogues, or interactive tasks).
3. **Formal rule systems** (logical, legal, or domain-specific constraints).

For each benchmark we ask:

- How is context represented, varied, or ignored?
- What is objectively verifiable versus human-judged?
- Do tests surface *interference* when norms and instructions collide, or measure *normality shifts* when a narrative frames an unusual practice as standard?

Connection back to individual control.

Benchmarks are technical artefacts, but they also embed a regulatory theory: they operationalise *whose interests count* when we score a model. The current crop still centres a single user prompt or a single row of text, thereby inheriting the individual-control focus critiqued in the previous chapters. Chapter 6 shows that, without context-rich evaluation, even the best alignment scores preserve only the appearance of individual autonomy while letting relational and systemic harms slip through. In other words, the benchmark itself becomes a ‘tech fix’ — effective for row-level guarantees yet blind to collective impact. By exposing that blind spot, the chapter completes the dissertation’s arc from collection to erasure to utilisation, demonstrating that any future remedy must pair individual controls with context-sensitive testing and collective-impact metrics if it is to achieve meaningful accountability.

6.2 Introduction

Machine-learning governance increasingly hinges on a model’s ability to *behave appropriately in context*. The two prior empirical chapters demonstrated why: synthetic data can mask lineage and legitimate biased pipelines; machine unlearning, while empowering individuals, still leaves participants anxious about downstream social effects. Both tools — insofar as they operate at the level of *data control* — treat harm as something that can be fixed once the ‘right’ records are present or absent. Yet real-world harms often materialise *after* training, when an LLM must interpret instructions, recognise tacit social cues, or follow organisational policy. Put bluntly, a privacy-perfect dataset is useless if the deployed model cannot tell when ‘joking’ about medical advice or leaking private emails is out of bounds.

Even a perfectly ‘de-identified’ training set and a fully compliant unlearning API leave people exposed if, at inference time, a model cannot tell that repeating a private e-mail or medical snippet is out of bounds for *this* requester in *this* setting. Hence the final link in the dissertation’s chain is to ask whether today’s evaluation culture can actually detect such contextual privacy breaches before systems reach end users.

This chapter therefore pivots from data pipelines to **behavioural evaluation**. We ask: *How do today’s benchmarks measure whether a large language model (LLM) adheres to explicit instructions, implicit contextual norms, and formal rule systems — especially when those constraints collide?* Answering requires crossing disciplinary lines. Work in NLP and HCI has produced dozens of specialised tests: verifiable-instruction suites (e.g. IFEval), cultural-norm vignettes (NormAd), privacy judgements grounded in Contextual Integrity (PrivaCI-Bench), and logic-puzzle datasets that read like textbook proofs (RuleBench, ProofWriter). Yet no survey has stitched these strands together to reveal where they complement — or contradict — one another.

Our review proceeds in three conceptual layers that mirror the escalating demands placed on deployed assistants and agents:

1. **Instruction fidelity** — can the model satisfy concrete, programme-checkable constraints on style, content, format, or length?
2. **Contextual norm sensitivity** — does it recognise what counts as polite, safe, or fair behaviour for *this* speaker, audience, and setting?
3. **Rule-based reasoning** — when supplied with explicit logical or legal rules, can it trigger the right rule and execute the inference correctly, even if the rule contradicts its pre-trained ‘intuition’?

Across these layers we uncover a persistent methodological tension: the more objectively verifiable a benchmark becomes, the more it drifts toward synthetic, toy-like constraints; the richer the social scenario, the heavier its reliance on fallible LLM-or human-judges. We also find two blind spots directly relevant to the dissertation’s through-line on relational harms: (i) almost no benchmark systematically manipulates the perceived ‘normality’ of a norm to

test whether models over-generalise dominant-culture expectations, and (ii) very few isolate the *interference* cost when explicit instructions and contextual norms clash.

The contributions of this chapter are as follows:

- We provide the *first comparative map* of 30+ adherence benchmarks spanning instructions, norms, rules, long-context processing, and safety jail-break tests.
- We introduce a five-axis coding scheme — constraint type, context representation, verification method, interference design, norm-normality manipulation — that enables gap-oriented synthesis.
- We surface concrete research opportunities, including controlled interference suites and partially script-verifiable norm evaluations, which Chapter 7 explores further.

The next section (Section 6.3) details the related work, followed by our literature-collection protocol (Section 6.4); Sections 6.5 - 6.8 then analyse the state of the art along the three layers above, culminating in a synthesis of open problems (Section 6.9) and a roadmap for future work (Sections 6.10).

6.3 Related Work

Machine learning evaluation is now a research field in its own right, yet only a fraction of the literature tackles the question central to this chapter: *Can we test whether a language model behaves appropriately when the ‘right’ answer depends on social situation rather than on surface-text cues?* We synthesize three streams of work that, together, define the state of the art and reveal the blind spots that motivate our benchmark analysis.

Context in NLP Evaluation

The earliest LLM leaderboards — SQuAD, GLUE, and SuperGLUE — treated ‘context’ as nothing more than a fixed block of text that feeds facts to the model. Critics quickly showed how brittle that assumption is: change the way the very same passage is written and the headline scores collapse. META-Eval, for instance, rewrites one SQuAD paragraph in three styles — encyclopedic Wikipedia prose, slick press-release copy, and informal conversational notes. Although every version conveys identical facts, models lose as much as 25 F1¹ on the chatty, vernacular rendering [109]. Early work outside the generative-LM boom already sounded this alarm: Cai et al. showed that even fine-tuned BERT-style classifiers lose dozens of F1 points when the same input is re-sliced by genre or demographic, and proposed “adaptive slicing” to surface those hidden shifts [45]; our review finds that most LLM adherence suites still lack an equivalent mechanism. Such sensitivity motivated

¹The F1 score is the harmonic mean of precision and recall; in question-answering it captures how completely and accurately a model extracts the correct span.

richer dataset documentation: Bender and Friedman’s ‘data statements’ and Gebru et al.’s ‘datasheets’ argue that without capturing speaker demographics, collection incentives, and power relations, benchmark numbers mislead more than they inform [26, 110].

Prompt-engineering studies extend the point. Gan and Mori et al. show that changing a single modifier — “write *succinctly*” versus “write *briefly*” — can swing factual accuracy by ten percentage points [107]. Similar fragility appears in summarisation: Borhan and Bajaj show that swapping a plain “Summarise the article in 100 words” prompt for a role-play framing (“You are a media-relations officer drafting a press release . . .”) slashes ROUGE-F by approximately 14 points and halves BLEU precision across both GPT-3.5 and Llama-2, underscoring how minor prompt wording can erode coherence and coverage [35]. Together these findings make clear that *context representation* — genre, register, communicative role — is a first-order variable, not a wrapper. Chapter 6 therefore treats context modelling as a core dimension when mapping LLM adherence benchmarks.

A second wave of work broadens the definition of context beyond linguistic style. Perez et al. embed innocuous biographical details in the system prompt and show that downstream moral judgements flip signs [89]; Long Context benchmarks reveal that evidence placed far from the query is often ignored, leading to shallow, order-dependent reasoning. Despite these advances, most studies still cast ‘context’ as text to be parsed rather than as the social norms that regulate who may transmit what to whom.

Connection to dissertation: the evidence above establishes that models are brittle even under benign linguistic restyling; in the following discussion, we extend the critique to *norm-governed* contexts where roles, obligations, and power differentials — not wording — define correctness.

Instruction Tuning, Alignment, and Their Limits

Instruction-tuned models such as FLAN-T5 and Llama 2 achieve broad-spectrum helpfulness by combining supervised fine-tuning with Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimisation [59, 200, 215, 259]. Constitutional AI layers a self-critique loop that references a fixed list of principles (e.g. ‘no medical advice’) to further reduce overt toxicity [18]. These pipelines succeed on headline metrics, yet empirical audits unearth lingering blind spots: models still hallucinate citations when retrieval contexts contradict world knowledge [184], over-refuse safe instructions that merely *look* disallowed [293], and default to Western etiquette in cross-cultural dialogue [220]. Alignment, in short, is uneven and context-dependent.

Recent proposals attempt to patch these weaknesses with task-specific fine-tunes or guardrails, but the fixes remain ad hoc. Larger, more diverse instruction corpora do raise average benchmark scores, yet they do not grant the model an explicit *reasoning* mechanism for deciding which rule overrides another when norms collide. Nor do they furnish a diagnostic framework to measure such trade-offs during evaluation. Our chapter shows that existing benchmarks rarely probe these collision cases, leaving practitioners without evidence on whether alignment strategies generalise beyond the training distribution.

Connection to dissertation: understanding the failure modes of instruction-tuned systems is prerequisite to any technical or policy remedy; our survey therefore catalogues which benchmarks can (and cannot) surface those failures.

Normative Frameworks and Sociotechnical Harms

Legal and sociological scholarship reminds us that many algorithmic harms cannot be reduced to per-interaction privacy breaches. Contextual Integrity (CI) conceptualises privacy as conformance to norms of information flow that differ across social spheres [196]. CI now informs safety and privacy test suites such as CASE-Bench and PrivaCI-Bench, where sender, recipient, attribute, and transmission principle are varied systematically to gauge model judgement [252, 165]. In parallel, harm taxonomies map representational, allocative, quality-of-service, and systemic injuries that fall outside traditional privacy law [237]. Relational data governance frameworks underscore that one person’s disclosure affects the collective, challenging notice-and-consent defaults [263].

Yet translating these rich theoretical constructs into *verifiable* tests has proved difficult. Existing CI-inspired benchmarks focus narrowly on a few transmission principles and binary safe/unsafe labels; they do not capture the nuanced, often contested nature of social norms nor the collective dimension of relational harms. Moreover, no mainstream suite allows evaluators to toggle norm salience or to measure how norms interact with explicit user instructions.

Connection to dissertation: by charting where normative theory *is* and *is not* operationalised, we expose the measurement gaps that permit relational and contextual harms to persist unseen.

6.4 Methodology

Our protocol adapts systematic-review practices to the fast-moving LLM literature but intentionally keeps the coding frame light, recognising that benchmark papers are heterogeneous.

Search strategy

We queried ACL Anthology, arXiv, Semantic Scholar and Google Scholar (Jan 2020–Mar 2025) with Boolean strings such as:

```
(\instruction following" OR \constraint following") AND (benchmark OR dataset)
AND (LLM)
(\social norm" OR \cultural norm") AND (evaluation OR benchmark)
(\rule following" OR \logical reasoning") AND (RuleTaker OR RuleBench)
```

Eligibility

Papers were included if they

- reported empirical results on a model >1 billion parameters,
- introduced *or* reused a benchmark whose *primary* aim is to test instruction, norm, or rule adherence,
- embedded those constraints in some textual context (prompt, dialogue, narrative, agentic environment), and
- provided at least one quantitative metric.

The final corpus contains 33 benchmarks.

Analytic frame

Each paper was coded along four categories:

1. **Constraint Type:** instruction, norm, or rule.
2. **Context Representation:** static prompt, dialogue, CI parameters, simulated web task, long document, etc.
3. **Verification Method:** deterministic script, symbolic-logic checker, LLM-judge, human panel, or hybrid.
4. **Collision Design:** none, implicit (observed but not controlled), explicit (benchmark deliberately engineers conflict).

Coding workflow and reliability.

All coding was done by the author in two passes, separated by a one-week wash-out. Pass 1 populated the four core fields and a free-text memo of edge-cases. Pass 2 re-coded a stratified 25 % sample (8 instruction, 6 norm, 3 rule benchmarks). Agreement with the original codes was 92 % (46/50 items). Discrepancies were resolved by memo-review and, where necessary, updating the category definitions.

Because the benchmarks span very different goals and grain sizes, we use a *narrative* rather than tabular synthesis. Sections 6.5, 6.3 – 6.3, and 6.8 proceed in four steps. First, three core sections — one each for *explicit instructions*, *contextual norms*, and *formal rules* — contrast emblematic datasets, showcase methodological innovations, and surface recurrent blind-spots inside their own constraint type. Section 6.8 then zooms out to the *meta-variables of context*: narrative framing, long-context windows, and alignment/safety pressure, all of which modulate adherence across constraint types. Finally, Section 6.9 braids the evidence

from all four analytic blocks into three cross-cutting methodological gaps — context representation, collision (interference) testing, and norm-normality manipulation — which seed the forward-looking proposals developed in the dissertation’s general discussion.

Limitations

- **Rapid-growth bias:** Papers published after March 2025 are not captured; a living-review appendix is planned.
- **LLM-judge reliance:** Several benchmarks use GPT-4-style models as evaluators; our coding distinguishes this but cannot fully quantify cascading error if those judges share blind spots with the evaluated systems.
- **English-centrism:** Despite targeted searches, non-English norm benchmarks remain under-represented — mirroring a gap the review itself problematises.

6.5 Benchmarking Instructions in Context

Overview of Instruction Following

The ability of LLMs to follow natural language instructions is a cornerstone capability, directly influencing their alignment with human intentions and their overall safety and reliability. Instruction following enables LLMs to perform desired tasks, generate appropriate responses, and operate within specified constraints. While advancements in instruction tuning through methods like Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO), often utilizing large-scale annotated or synthetic datasets, have significantly improved general instruction-following capabilities, numerous studies indicate that current LLMs still struggle, particularly with complex instructions involving multiple, fine-grained, or conflicting constraints. This deficiency limits their applicability in real-world scenarios that demand precise adherence to intricate requirements [137]. Evaluating this capability accurately and robustly has thus become a critical research problem [273]. Baseline tests of atomic, script-checkable constraints reveal how brittle instruction-following remains before social context is even introduced. But the very ease of automatic verification can be misleading: a system may nail a comma-count requirement while flouting the substantive constraints that matter for real-world safety and fairness.

Key Benchmarks and Methodologies

Several benchmarks have been developed to specifically evaluate LLM instruction-following abilities, each with distinct methodologies and focuses.

- IFEval (Instruction-Following Eval):** IFEval represents a significant effort towards objective and reproducible evaluation by focusing on ‘verifiable instructions’ [292]. These are instructions whose fulfilment can be automatically checked using programmatic methods, typically Python scripts [102]. The benchmark includes 25 distinct types of verifiable instructions covering aspects like keyword usage (inclusion, frequency, exclusion), response language, length constraints (words, sentences, paragraphs), detectable content (postscripts, placeholders), formatting (bullets, titles, JSON), case changes (uppercase, lowercase), start/end requirements, and punctuation rules (e.g., no commas). Approximately 500 prompts were constructed, each containing one or more such instructions. Adherence is measured using strict accuracy (requiring all instructions in a prompt to be followed perfectly) and loose accuracy (allowing for minor variations by applying transformations to the response before checking). The primary strength of IFEval lies in its high verifiability and objectivity stemming from automated script-based validation. However, this focus on simple, verifiable constraints means it may not fully capture the complexity of constraint interactions or composition [191]. Furthermore, its instructions are typically presented within the prompt itself, offering limited assessment of adherence within richer narrative contexts, and it does not directly evaluate norm adherence or interference effects. The synthetic nature of some verifiable constraints might also limit the direct transferability of findings to all real-world instruction types [101]. Despite these limitations, IFEval provides a valuable foundation for objective adherence measurement, potentially adaptable to evaluating adherence to verifiable aspects of norms or rules.
- ComplexBench:** Addressing the limitation of benchmarks focusing on single constraints, ComplexBench specifically targets the evaluation of complex instructions composed of *multiple* interacting constraints [273]. It introduces a hierarchical taxonomy of constraints (including 4 constraint types and 19 dimensions) and 4 composition types, along with dependency graphs to model constraint interactions. The benchmark comprises 1,150 manually constructed instructions, ensuring high quality but limiting scalability. Evaluation employs a novel method called Rule-Augmented LLM-based (RAL) evaluation, which combines rule-based checks with LLM judgments, reportedly achieving better correlation with human assessment than standard LLM-as-judge approaches. The final score considers the dependency structure defined by the constraint composition. Metrics include assessing whether all scoring questions within an instruction are correct (Original Test) and whether groups of related tasks are correct (Coherent Test). ComplexBench’s strength lies in its explicit focus on constraint composition, reflecting a crucial aspect of real-world instruction complexity. However, its reliance on manual construction and partial LLM judgment introduces limitations in scale and potential evaluator bias. Like IFEval, its focus remains on explicit instructions rather than implicit contextual norms or interference. ComplexBench represents a step towards evaluating adherence in more realistic, multi-faceted instruction scenarios.

- FollowBench:** FollowBench evaluates adherence to multi-level, fine-grained constraints, categorized into five types: Content, Situation, Style, Format, and Example [150]. Its core methodological innovation is a multi-level mechanism where constraints are added incrementally to an initial instruction, level by level. This allows for precise measurement of how LLM performance degrades as instruction complexity increases. Evaluation involves prompting strong LLMs (e.g., GPT-4) with ‘constraint-evolution paths’ to assess satisfaction of individual constraints in potentially open-ended responses. Key metrics include Hard Satisfaction Rate (HSR: all constraints satisfied), Soft Satisfaction Rate (SSR: average satisfaction of individual constraints), and Consistent Satisfaction Levels (CSL: number of consecutive levels passed starting from level 1). The evaluation combines rule-based checks with these LLM-based assessments. FollowBench offers a systematic way to assess adherence to different *types* of constraints and resilience to increasing complexity. However, the LLM-based evaluation component introduces potential subjectivity. While the ‘Situation’ constraint type hints at context, the benchmark primarily evaluates adherence to explicit constraints rather than broader contextual norms or interference effects. FollowBench contributes a structured approach to understanding how different constraint types and cumulative complexity impact LLM performance.
- FoFo (Format Following):** FoFo is a benchmark specifically designed to evaluate LLMs’ ability to follow complex, domain-specific formats, a capability crucial for AI agents but often overlooked in general instruction-following benchmarks [283]. It features a diverse range of real-world formats (e.g., HL7-CDA in healthcare) and instructions developed through an AI-Human collaborative method. FoFo’s evaluation highlights that format-following performance can be independent of content generation quality and that open-source models often lag behind closed-source ones in format adherence.
- CELLO (Complex Instruction Evaluation):** CELLO evaluates LLMs’ capability to comprehend and follow complex instructions, addressing limitations in existing benchmarks that often use simple or close-ended tasks [133]. It defines complexity based on eight features, covering both complex task descriptions (e.g., multiple tasks, constraints) and complex inputs (e.g., long context, noise, multi-turn history). CELLO constructs its evaluation dataset from real-world scenarios across nine tasks and proposes four evaluation criteria with corresponding metrics to systematically assess performance.
- Other Related Benchmarks:** Other notable benchmarks include **InFoBench**, which uses a ‘decomposed requirements following ratio’ (DRFR) metric evaluated by LLMs; **WILDIFEVAL**, a large-scale (12K) dataset focusing on real user instructions with diverse, multi-constraint conditions categorized into eight classes [170]; **VFF (Verifiable Format Following)**, a dataset focused on fine-grained, fully verifiable format following, used for progressively training smaller LLMs [269]; and **CFBench (Com-**

prehensive Constraints Following Benchmark), a large-scale (1,000 samples) benchmark covering real-life scenarios and NLP tasks, emphasizing multi-dimensional assessment criteria and requirement prioritization [290]

- These further illustrate the community’s efforts to move beyond simple instruction following towards more nuanced evaluations, including real-world complexity and specific constraint types like formatting.

Findings show that LLM fragility at this ‘clean-room’ level sets an upper bound on later, more contextual performance.

Handling of Context in Instruction Following

The reviewed instruction-following benchmarks primarily handle context in a limited way. The ‘context’ is typically the instruction prompt itself, which contains the task description and the explicit constraints the LLM must follow. Some benchmarks, like FollowBench with its ‘Situation’ constraint type [150], might embed minimal situational details within the prompt. However, none of these benchmarks systematically evaluate instruction following when the instruction is embedded within a rich, dynamic narrative or a complex, evolving situational context external to the prompt itself. The context is generally static and provided upfront. The almost total absence of narrative setting here underscores a blind spot for group-level and relational effects.

Verifiability and Objectivity

A significant distinction among these benchmarks lies in their approach to verifiability and objectivity. IFEval stands out for its use of automated scripts, ensuring high objectivity and reproducibility for its specific set of verifiable constraints. This aligns well with the need for rigorous, objective evaluation methods. In contrast, ComplexBench (using RAL evaluation) and FollowBench (using LLM-as-judge) incorporate LLM judgments to handle more complex or open-ended instructions that are difficult to verify purely programmatically. While this allows for evaluating adherence to more nuanced instructions, it introduces potential evaluator biases, inconsistencies, and robustness issues associated with the LLM-as-judge paradigm. This highlights a fundamental trade-off in current evaluation methodologies: prioritizing strict, objective verifiability often restricts evaluation to simpler constraints, while assessing complex, realistic instructions often necessitates reliance on less objective LLM or human judgments. The trade-off between rigid objectivity and real-world realism foreshadows practical enforcement dilemmas.

Insights and Gaps (Instruction Following)

The development trajectory of instruction-following benchmarks reveals a clear progression. Initial efforts like IFEval prioritized objective verifiability through quantifiable, often

atomic constraints [292]

Recognizing that LLMs often fail when constraints are combined, subsequent benchmarks like ComplexBench focused explicitly on evaluating the composition and interaction of multiple constraints, employing hybrid evaluation methods like RAL. FollowBench introduced a graded approach, measuring performance degradation as constraints are incrementally added, providing insights into model robustness against increasing complexity. Benchmarks like FoFo and CELLO further refined the evaluation of specific aspects like format following and handling complex instruction features. This evolution reflects an increasing awareness that real-world instruction following involves handling complex, multi-faceted requirements, pushing evaluation beyond simple checks.

This progression also underscores an inherent tension in evaluation design. The rigorous objectivity of script-based methods like IFEval comes at the cost of potentially evaluating overly simplistic or synthetic instructions. Conversely, benchmarks like FollowBench and ComplexBench tackle more realistic complexity but introduce the potential subjectivity and biases of LLM-based evaluation. Finding the right balance between evaluation rigour and task realism remains a central challenge.

Despite these advancements, a significant gap persists: the evaluation of instruction following within truly rich, dynamic narrative or situational contexts. Current benchmarks largely confine instructions to static prompts. Furthermore, while complexity is addressed through composition (ComplexBench) and accumulation (FollowBench), the specific *interference* effects between different types of constraints, or between explicit instructions and implicit contextual factors, are not systematically isolated or measured. Current benchmarks seldom probe what happens when explicit instructions collide with external social demands.

6.6 Benchmarking Contextual Norms

Defining Contextual Norms

Contextual norms encompass the social, cultural, ethical, and situational rules that govern appropriate behaviour within a specific context [166]. These norms differ from explicit instructions or formal logical rules in that they are often implicit, learned through social interaction, culturally variable, and highly dependent on the specific situation (e.g., sender, receiver, setting). For LLMs intended for global deployment, understanding and adhering to diverse cultural norms is crucial to avoid generating offensive, inappropriate, or biased content [220]. However, LLMs often reflect the biases present in their training data, frequently exhibiting a slant towards dominant Western or English-centric cultural perspectives and struggling with nuances from under-represented cultures. Evaluating an LLM’s ability to perceive and adhere to these contextual norms is therefore essential for building inclusive and trustworthy AI systems. Shifting focus from private data to shared social expectations foregrounds relational dimensions of harm.

Benchmarks for Cultural and Social Norms

Several benchmarks have emerged to evaluate LLMs’ understanding and adherence to cultural and social norms, moving beyond simple instruction following.

- CulturalBench:** This benchmark focuses on assessing LLMs’ explicit *cultural knowledge* [55]. It contains 1,227 questions covering 17 diverse topics (e.g., food, greetings, etiquette) across 45 global regions, including under-represented areas. The questions are human-written and rigorously verified by five independent annotators each, ensuring high quality. Evaluation occurs in two modes: CulturalBench-Easy (standard multiple-choice) and CulturalBench-Hard (converting each option into a binary True/False question). The ‘Hard’ setup proved significantly more challenging for LLMs, revealing sensitivity to question framing and substantial gaps compared to human performance. The benchmark also includes questions with multiple correct answers, highlighting LLMs’ tendency to converge on a single option. While valuable for assessing foundational knowledge, CulturalBench primarily tests recall via QA rather than behavioural adherence within a narrative or interactive context, and it does not directly assess interference or manipulate norm expectations. Its dataset size is also relatively small. It serves as a measure of the prerequisite knowledge needed for culturally sensitive behaviour.
- NormAd:** NormAd shifts the focus from knowledge recall to *cultural adaptability*, measuring how well LLMs judge the social acceptability of actions described in narrative stories [220]. It includes 2.6k stories operationalizing social etiquette norms (used as a proxy for culture) from 75 countries, based on the Cultural Atlas resource. A key feature is its representation of norms at three levels of granularity: explicit Rule-of-Thumb, abstract Value, and implicit Country context. The task requires the LLM to judge the social acceptability (Yes/No/Neutral) of the behaviour depicted in the story, given one of these levels of contextual information. Evaluation uses accuracy compared to human judgments. NormAd demonstrates that current LLMs struggle with cultural reasoning across all context granularities, performing better with explicit rules but still lagging behind humans, and exhibiting biases towards English-centric cultures and agreement (finding it easier to judge norm-adherent stories). By embedding norms within narrative vignettes and testing adaptability across abstraction levels, NormAd moves closer to evaluating contextual understanding. However, it focuses on *judgment* rather than *generation* of norm-adherent behaviour and doesn’t explicitly test interference or normality manipulation. It provides valuable insights into LLMs’ socio-cultural reasoning pitfalls [252].
- NormDial:** This benchmark introduces a bilingual (Chinese/English) dataset of dyadic dialogues specifically designed for modelling social norm adherence and violation within conversations [166]. It features turn-by-turn annotations indicating whether a speaker’s utterance adheres to or violates relevant social norms. The associated task is so-

cial norm observance *detection*. The dataset was synthetically generated using LLMs prompted with expert-annotated norms, followed by human-in-the-loop quality control. NormDial’s strength lies in its focus on evaluating norm understanding within the dynamic context of dialogue, offering bilingual comparability. However, its reliance on synthetic data may not capture all real-world complexities, and the primary task is detection, not generation or adherence. It doesn’t explicitly test interference or normality manipulation. NormDial represents an important step towards evaluating norms in interactive settings.

- **Casa (Cultural and Social Awareness benchmark):** Casa evaluates LLM *agents*’ sensitivity to cultural and social norms within simulated web-based tasks: online shopping and social discussion forums [214]. It leverages established cultural taxonomies across 17 countries to create scenarios [214]. Two main scenarios are used: S1 tests sensitivity to norm violations within user queries (distinguishing between queries that adhere vs. violate norms), and S2 tests sensitivity to norm-violating content observed on simulated websites [214]. Evaluation uses an LLM judge (GPT-4o) and specific metrics: Helpfulness Rate (Help-R for S1-Adhere), Awareness Coverage Rate (AC-R for S1-Violate), Educational Rate (Edu-R for S1-Violate), and Violation Rate (Vio-R for S2, measuring reliance on misleading web content) [214]. Casa uniquely evaluates norm adherence in an agentic setting involving interaction with external web context. Its findings indicate that LLM agents struggle significantly more in these web-based agent settings compared to non-agent environments. Limitations include the reliance on LLM-as-judge evaluation. It focuses on detecting/responding to violations or adhering to norms, rather than explicitly testing interference or normality manipulation. Casa provides a valuable framework for assessing norm adherence in applied, context-rich agentic tasks [252]
- **CASE-Bench (Context-Aware Safety Evaluation):** While primarily a safety benchmark, CASE-Bench offers a rigorous methodology for incorporating context into evaluations, drawing heavily on Contextual Integrity (CI) theory [252]. CI theory defines norms governing the appropriateness of information flow based on context parameters: sender, recipient, information subject, attributes (information type), and transmission principle (purpose, constraints). CASE-Bench applies this by creating pairs of contexts (one intended safer, one less safe) for 450 controversial/harmful queries across 45 categories, resulting in 900 query-context pairs. Contexts are formally described using CI parameters. The benchmark relies on large-scale human annotations (over 2000 annotators, 21 per item based on power analysis) to establish ground truth regarding whether responding to a query is safe within a given context. The core task evaluates whether an LLM’s *safety judgment* (safe/unsafe to respond) aligns with the aggregated human judgment for a given query-context pair. CASE-Bench demonstrates the substantial and statistically significant influence of context on safety judgments and reveals issues like over-refusal by commercial models in safe contexts. Its strengths include

the rigorous application of the transmission components CI theory and high-quality human annotation. However, it evaluates safety *judgment* rather than response generation, and doesn't explicitly account for norms - it doesn't directly test interference or normality manipulation but provides a strong model for context formalization.

Taken together, these five resources chart real progress, yet they also underscore a common blind spot. Each benchmark broadens cultural coverage relative to early English-centric tests, but the union of them still leaves large regions, languages, and minority sub-cultures unrepresented, and they rarely place contrasting norms *in direct tension* with an explicit instruction. As a result, a model can score well and still fail when (i) a request comes from a community missing in the dataset, (ii) a local expectation collides with a global instruction, or (iii) the narrative frames an otherwise 'abnormal' practice as customary. In short, advertising an LLM as 'context-aware' on the basis of today's etiquette benchmarks may amount to premature certification unless future evaluations close three methodological gaps developed later in this chapter: *representational breadth*, *interference stress-tests*, and *normality manipulation*.

Handling of Narrative Context

Benchmarks assessing norm adherence show a greater emphasis on narrative and situational context compared to instruction-following benchmarks. NormAd uses narrative stories, NormDial employs dyadic dialogues, and Casa situates LLM agents within web browsing scenarios. CASE-Bench defines context through formalized CI parameters within specific scenarios. This contrasts sharply with CulturalBench, which tests cultural knowledge using decontextualized question-answering. This reflects the understanding that norms are inherently contextual and must be evaluated within relevant situational or narrative frames. Embedding norms in stories demonstrates that appropriateness judgments change as the surrounding narrative shifts.

Verifiability and Objectivity

Evaluating adherence to social and cultural norms presents significant challenges for objective verification due to the inherent subjectivity and cultural relativity of norms themselves. CulturalBench relies on standard QA accuracy against human-verified answers. NormAd compares LLM judgments against human consensus on social acceptability within stories. NormDial uses turn-level annotations, which could be human or LLM-verified. Casa utilizes LLM-as-judge metrics. CASE-Bench employs large-scale human annotation as its ground truth, offering high reliability but at considerable cost. None achieve the purely automated verifiability of IFEval for simple instructions. The field relies heavily on human consensus or LLM proxies to establish ground truth for norm adherence, highlighting the difficulty in creating fully objective evaluations for these complex, subjective phenomena.

Heavy reliance on human or LLM judges highlights how evidence becomes murkier once provenance and consent blur.

Insights and Gaps (Norm Adherence)

The evolution of norm adherence benchmarks indicates a necessary shift from testing static knowledge (like CulturalBench) towards evaluating judgment and behaviour within specific, dynamic contexts. This includes assessing responses to narrative scenarios (NormAd), interactions in dialogues (NormDial), actions in agentic tasks (Casa), and judgments within formally defined contexts (CASE-Bench). This progression acknowledges that simply knowing a norm is insufficient; models must demonstrate the ability to apply that knowledge appropriately in context.

Frameworks like Contextual Integrity (CI), used in CASE-Bench and PrivaCI-Bench [165], offer a structured way to define and evaluate context, particularly for norms related to privacy and safety. By parametrizing context (sender, receiver, etc.), these benchmarks enable systematic variation and measurement of its impact. This enhances rigour but may struggle to encompass the full ambiguity and breadth of general social or cultural norms, which often extend beyond information flow principles.

Across these diverse benchmarks, a consistent pattern emerges: LLMs exhibit significant limitations in handling normative complexity. They show biases towards dominant cultures (NormAd), sensitivity to question framing (CulturalBench), agreement bias (NormAd), degraded performance in agentic contexts (Casa), and over-refusal in safe contexts (CASE-Bench). These findings underscore the fragility of current LLMs' normative reasoning and the critical need for robust, diverse evaluation methods.

However, significant gaps remain. Firstly, there is a lack of benchmarks designed to explicitly test the *interference* between contextual norms and explicit instructions when they conflict. Secondly, existing methods do not systematically manipulate the *perceived normality* or expectedness of a contextual norm within a narrative to evaluate how LLMs handle deviations from their learned statistical patterns of social behaviour. Few evaluations yet examine direct conflicts between what the prompt commands and what the setting deems acceptable.

6.7 Benchmarking Rules and Logical Constraints

Rule Following vs. Instruction Following

While related, rule-following, in the context of benchmarks like RuleBench, RuleTaker, and ProofWriter, is typically distinguished from general instruction-following. Instruction-following often pertains to executing behavioural commands or adhering to constraints on

output format, style, or content. Rule-following, conversely, emphasizes adherence to explicit, often formal or semi-formal, *inferential* rules that govern logical reasoning [253]. These rules are frequently abstract (e.g., using variables like ‘X’ or ‘person A’), conditional (‘if P then Q’), and require instantiation based on specific facts or context to derive conclusions. Evaluating this capability is crucial, particularly for applications demanding high levels of logical consistency, faithfulness, and reliability, such as legal analysis, scientific reasoning, or complex problem-solving. Current LLMs, despite advancements, often struggle with complex multi-step logical reasoning, consistency, and avoiding fallacies. Simply stating a rule is no guarantee a model will notice or implement it faithfully.

Key Benchmarks and Methodologies

Several benchmarks specifically target the evaluation of logical rule-following.

- RuleBench:** This benchmark is explicitly designed to evaluate *inferential rule-following*, distinguishing it from instruction-following [253]. It assesses LLMs across various reasoning tasks (relation extraction, QA, math, legal judgment) by providing a context/question and a set of explicit inferential rules [253]. A key aspect is evaluating the LLM’s ability to both *trigger* the correct rule (identify the relevant one, potentially amidst noise rules) and *execute* it accurately to derive the answer. RuleBench systematically varies evaluation conditions, including the number of rules provided (golden rule only, few rules with noise, all rules with more noise, no rules), the form of the rules (natural language vs. formal language/FOL), the use of counterfactual rules (where the rule’s conclusion contradicts known facts), and the application of Chain-of-Thought (CoT) prompting. Evaluation metrics include overall accuracy and, importantly, granular error types: Triggering Error and Execution Error. RuleBench provides a rigorous framework for dissecting inferential rule-following capabilities and identifying specific failure modes. Its limitations include the focus on explicit rules derived from existing datasets and the limited exploration of normality manipulation beyond logical counterfactuals. It does not explicitly test interference with behavioural instructions.
- ProofWriter:** ProofWriter adopts a generative approach, using a T5 model to not only determine the truth value of a hypothesis given an NL theory but also to *generate* the natural language proof supporting the conclusion [255]. It can also generate logical implications of a theory and perform a form of single-fact abduction (identifying a missing fact needed to complete a proof). A key technique is iterative 1-step implication generation: the model repeatedly generates a direct (1-hop) inference and its simple proof, adds the inference back to the context, and continues until no new implications are found. Multi-step proofs are assembled from these 1-hop fragments. Evaluation focuses on the accuracy and faithfulness of the generated proofs, as well as the correctness of generated implications and facts. ProofWriter claims its generated proofs

are faithful representations of the model’s reasoning process, achieving state-of-the-art results on RuleTaker [227]. Its strength lies in producing verifiable and interpretable proof outputs. However, it still operates on explicit NL rules and facts and does not address social norms or context beyond the provided theory.

Handling of Context and Verifiability

In these rule-following benchmarks, the ‘context’ is almost exclusively defined by the explicit set of facts and rules provided as input alongside the query or hypothesis. The task is to reason strictly within this closed world. Verifiability is generally high. RuleBench uses accuracy supplemented by error analysis based on whether the correct rule was triggered and executed. ProofWriter allows for verification of the generated proofs themselves, either against gold proofs or through logical checkers, enhancing transparency. Rule-following studies operate in closed worlds, leaving open questions about real-world, open-context robustness.

Insights and Gaps (Rule Following)

The design of RuleBench highlights the value of decomposing rule-following failures into distinct stages: *triggering* (identifying the correct rule) and *execution* (applying it correctly). This granular analysis reveals that LLMs can falter at either step — failing to recognize a rule’s relevance even if they understand it, or misapplying a rule they correctly identified. This provides more targeted insights for improvement than simply measuring end-to-end accuracy.

Furthermore, consistent findings across benchmarks like RuleBench and studies analysing logical reasoning [202] show that LLMs struggle significantly more when rules are counterfactual (conflicting with pre-trained knowledge) or involve logical complexities like negation. This suggests that LLMs are not purely abstract rule followers but heavily rely on patterns and knowledge learned during pre-training. Adherence diminishes when explicit rules clash with this internal model or demand sophisticated logical manipulation, indicating a blend of statistical pattern matching and nascent symbolic reasoning capabilities.

While these benchmarks rigorously evaluate adherence to explicit, often formalisable rules, they do not directly address the challenge of adhering to implicit, fuzzier social norms often encountered in real world contexts, and especially in contexts that are further out-of-distribution. The context is typically limited to the provided logical theory. Additionally, the potential interference between explicit rules and other contextual factors (like social norms or behavioural instructions) remains largely unexplored within these frameworks.

6.8 Benchmarking Context, Framing, and Alignment

Having surveyed the state of the art for *each* constraint type individually — explicit instructions (Section 6.5), social and cultural norms (Section 6.3 – 6.3), and formal rules (Section 6.7) — we now zoom out to ask what happens *across* those silos once the wider interaction setting comes into play. Empirical papers repeatedly show that a model judged faithful to an instruction, safe with respect to a norm, or logically sound under a rule can still behave unpredictably when the same task is wrapped in a different story, when the context window stretches from hundreds to tens of thousands of tokens, or when a safety-alignment objective overrides earlier fine-tuning. This section therefore tracks three cross-cutting variables that re-appear throughout the literature yet are seldom evaluated systematically:

- **Narrative framing** — how the story or role assignment surrounding a request shifts the model’s behaviour;
- **Context length & salience** — whether relevant evidence remains accessible as prompts grow and retrieval becomes noisier;
- **Alignment and safety objectives** — the guard-rails and jailbreak dynamics that can trump both instructions and norms.

By examining these patterns we move from a catalogue of benchmarks to an integrated account of *why* adherence breaks down in practice and what properties future evaluations must capture.

Context Sensitivity and Framing Effects

LLM responses are demonstrably sensitive to how information and tasks are presented.

- **Framing the Game:** This benchmark directly investigates the impact of narrative context framing on LLM decision-making, using the Prisoner’s Dilemma as a case study [225]. By procedurally generating diverse narrative vignettes that share the same underlying game-theoretic structure (payoffs), the benchmark systematically varies the framing. The findings reveal significant variability in LLM choices (cooperate vs. defect) based purely on the narrative framing, even when the rational payoff structure remains constant. This suggests LLMs inherit human-like cognitive biases and sensitivities to framing from their training data, deviating from purely rational decision-making. This sensitivity complicates evaluation, as observed adherence (or lack thereof) to a norm or instruction might be influenced by the specific framing rather than the core principle itself [174]. While not explicitly framed as norm adherence, the benchmark demonstrates how context framing can implicitly alter the perceived goals or norms governing behaviour in a situation.
- **Long Context Benchmarks:** The ability to process and reason over long contexts is crucial for understanding narratives and complex situations where instructions or

norms might be embedded. Benchmarks like LongBench (v1 and v2)[255], LongGenBench [173], BABILong [162], and Loong [267] evaluate LLM capabilities on tasks requiring understanding of sequences ranging from tens of thousands to millions of tokens. Tasks include long-document QA, summarization, code understanding, and long-form generation [280]. A consistent finding across these benchmarks is that LLM performance often degrades significantly as context length increases. Models may struggle to retrieve or utilize information effectively across the entire context window, sometimes referred to as the ‘lost in the middle’ problem, effectively using only a fraction of the available context. Evaluating long-form *generation* quality presents distinct challenges compared to long-context *understanding*. While techniques like Retrieval-Augmented Generation (RAG) can help, they often do not fully compensate for limitations in core long-context processing. This performance degradation directly impacts the ability of LLMs to maintain consistent adherence to instructions or norms over extended interactions or within lengthy documents. LongGenBench, for instance, implicitly tests instruction adherence by requiring models to follow specific formatting and sequential answering instructions within the long generated text.

- **Contextual Integrity (CI) Evaluations:** Benchmarks like PrivaCI-Bench [165] and CASE-Bench [252] explicitly demonstrate the power of context in shaping judgments about appropriateness, particularly concerning privacy and safety norms. By applying CI theory to parametrize context (sender, recipient, subject, attributes, transmission principle), these benchmarks show that changing these contextual factors significantly alters whether an information flow (or responding to a query) is considered acceptable or safe, both by humans and (to varying degrees) by LLMs. They provide a formal methodology for incorporating and evaluating specific contextual elements relevant to norm adherence, but they do not attend particularly well to the role of norms.
- **Faithfulness Evaluation in Context:** Benchmarks like **FaithEval** specifically target the evaluation of LLM faithfulness to provided context, particularly in challenging scenarios relevant to RAG systems [184]. FaithEval includes tasks designed to simulate real-world retrieval issues: unanswerable contexts (where the answer isn’t present), inconsistent contexts (where different documents support different answers), and counterfactual contexts (where the context contradicts world knowledge). It comprises 4.9K problems validated through LLM and human checks. Findings indicate that even advanced models struggle with faithfulness, especially in these challenging contexts. Similarly, **ContextualBench** is mentioned as an evaluation framework compiling multiple RAG benchmarks with consistent settings to assess contextual faithfulness [194]. These benchmarks highlight the difficulty LLMs face in adhering strictly to provided context, especially when it is incomplete, contradictory, or counterfactual.
- **Evaluating Contextual Judgments: ContextualJudgeBench** addresses the evaluation of LLM-as-judge models specifically within *contextual* scenarios like RAG-QA and summarization, which are often overlooked by existing judge benchmarks focused

on non-contextual tasks like instruction following [284]. It includes 2,000 challenging response pairs across eight splits designed to reflect real-world contextual evaluation needs, such as assessing faithfulness, conciseness, or refusal appropriateness based on the provided context. The benchmark reveals that evaluating responses based on context presents a significant challenge even for state-of-the-art judge models.

Safety and Alignment Benchmarks

Safety and alignment evaluations often implicitly or explicitly test adherence to safety norms or principles, sometimes under adversarial conditions. Benchmarks like SafeDialBench [50], SG-Bench [186], and Agent-SafetyBench [291] are designed to assess LLM robustness against adversarial attacks aimed at circumventing safety protocols and eliciting harmful outputs. These often involve multi-turn dialogues or agentic scenarios to simulate more realistic interactions and employ diverse attack strategies (e.g., role-playing, purpose reversal, prefix injection). SafeDialBench, for instance, evaluates safety across multiple dimensions (Fairness, Legality, Morality, etc.) using a hierarchical taxonomy and assesses abilities like identifying risks, handling unsafe information appropriately, and maintaining consistency under attack within multi-turn dialogues. Failures on these benchmarks represent critical breakdowns in adherence to safety norms and rules, particularly under pressure or when faced with deceptive contexts. Over-refusal and jailbreak failures show that compliance mechanisms must extend beyond one-shot filtering.

Insights and Gaps (Context, Framing, Alignment)

The collective evidence from these diverse benchmarks strongly suggests that LLM adherence is not a fixed property but is highly malleable, influenced significantly by contextual factors. Narrative framing can alter strategic decisions even when underlying logic is constant (Framing the Game). Increasing context length often leads to performance degradation, impacting consistency (LongBench). Specific situational parameters, as formalized by CI theory, demonstrably shift judgments of safety and privacy appropriateness (CASE-Bench, PrivaCI-Bench). Benchmarks like FaithEval show that faithfulness to context breaks down when the context is unanswerable, inconsistent, or counterfactual. ContextualJudgeBench further reveals that evaluating contextual adherence itself is challenging for current LLM judges. This convergence underscores the necessity of context-aware evaluation; assessing adherence in a vacuum is insufficient and potentially misleading.

Furthermore, the evolution of safety and alignment research reflects this growing contextual awareness. Early safety evaluations often focused on single-turn interactions, whereas newer benchmarks like SafeDialBench emphasize multi-turn dialogues and adversarial robustness. This shift indicates a move towards evaluating and enforcing adherence not just as static filtering but as consistent, principle-based behaviour over extended, potentially challenging interactions.

Despite this progress, critical gaps remain. It is difficult to disentangle the specific impact of *contextual norms* from other confounding factors like general context sensitivity, narrative framing effects, or performance degradation due to context length. When a model fails in a long, normatively complex scenario, isolating the precise cause - failure to understand the norm, conflict with an instruction, framing influence, or simply cognitive overload due to length - is challenging with current methods. This opens up a risk that while technical process will be made along those lines, there's field blindness towards the role of norms. A persistent measurement gap: no standard method yet isolates how norms, instructions and safety rules interfere.

6.9 Synthesis of Methodological Gaps

The preceding review highlights significant progress in developing benchmarks to evaluate LLM adherence across various dimensions: instructions, norms, rules, and other forms of context. However, a synthesis across these areas reveals critical and recurring methodological gaps that hinder a comprehensive understanding of LLM behaviour in complex, context-rich scenarios.

Comparative Overview Table

To provide a structured overview of the landscape, Table 6.1 summarizes key characteristics of the representative benchmarks discussed.

Limitations in Manipulating and Evaluating Normality

A significant gap identified across the reviewed literature is the lack of systematic methods for manipulating and evaluating the *normality* or *expectedness* of contextual norms. Most benchmarks treat norms or rules as static elements within a given context. For instance, CulturalBench and NormAd assess knowledge or judgment related to established cultural norms, while RuleBench evaluates adherence to predefined logical rules. While RuleBench's counterfactual setting introduces rules that contradict factual knowledge, and FaithEval includes counterfactual contexts, these primarily test logical adherence versus factual recall or faithfulness to potentially false information, rather than manipulating the *perceived social typicality* of a behaviour within a specific narrative context.

There is a scarcity of benchmarks designed to explore how LLMs respond when a narrative explicitly frames an unusual or counter-intuitive behaviour as normal or expected *within that specific situation*. For example, how would an LLM respond in a story where, contrary to typical Western norms, interrupting a speaker is explicitly described as a sign of respect *in that particular fictional culture*? Understanding how LLMs reconcile such explicit contextual framing of abnormality with their vast pre-trained knowledge of statistical regularities

Figure 6.1: Comparative Overview of LLM Adherence Benchmarks

<i>Benchmark</i>	<i>Task Focus</i>	<i>Evaluation Method</i>	<i>Verifiability</i>	<i>Context Handling</i>	<i>Norm/Rule Representation</i>	<i>Interference Testing?</i>	<i>Normality Manipulation?</i>
<i>IFEval</i>	Verifiable Instruction Following	Automated Scripts	High	Static Prompt	Explicit Instructions (Keywords, Length, Format, etc.)	No	No
<i>ComplexBench</i>	Complex Instruction Following (Composition)	Rule-Augmented LLM Judge (RAL)	Medium (LLM Judge)	Static Prompt	Explicit Instructions (Taxonomy, Composition Types)	Implicit (Composition)	No
<i>FollowBench</i>	Fine-Grained Instruction Following (Levels)	LLM Judge + Rule-based	Medium (LLM Judge)	Static Prompt (incl. "Situation" constraint)	Explicit Instructions (Content, Situation, Style, etc.)	Implicit (Levels)	No
<i>FoFo</i>	Domain-Specific Format Following	AI-Human Collaborative	Medium	Static Prompt	Explicit Format Instructions (Domain-Specific)	No	No
<i>CELLO</i>	Complex Instruction Following (Features)	Custom Metrics (Faithfulness, Format, etc.)	Medium (Rule/LLM)	Complex Input (Long Context, Noise, Multi-turn)	Explicit Instructions (8 Features)	Implicit (Complexity)	No
<i>WILDIFEVAL</i>	Real-World Multi-Constraint Instructions	Decomposition + LLM/Human Eval	Medium	Real User Prompts	Explicit Constraints (8 Classes)	Implicit (Multiple Constraints)	No
<i>CulturalBench</i>	Cultural Knowledge (QA)	Accuracy (Multiple Choice / Binary)	High (vs. Verified QA)	Decontextualized QA	Implicit in QA (Topics: Food, Greetings, etc.)	No	Implicit (Easy/Hard Framing)
<i>NormAd</i>	Cultural Norm Adaptability (Judgment)	Accuracy (vs. Human Judgment)	Medium (Human Judg.)	Narrative Stories (Rule/Value/Country context)	Explicit (Rule-of-Thumb), Abstract (Value), Implicit (Country)	No	No
<i>NormDial</i>	Social Norm Detection (Dialogue)	Turn-level Annotation Check	Medium (Human/LLM)	Dyadic Dialogue Turns	Explicit (Annotated Norms)	No	No
<i>Casa</i>	Agent Norm Sensitivity (Web Tasks)	LLM Judge (AC-R, Edu-R, Help-R, Vio-R)	Medium (LLM Judge)	Simulated Web Context (Queries & Observations)	Implicit (Cultural/Social Norms for Countries)	No	No
<i>CASE-Bench</i>	Contextual Safety Judgment	Comparison to Human Judgments	High (vs. Human Ann.)	CI-Formalized Contexts (Safe/Unsafe Scenarios)	Explicit (CI Norms: Sender, Receiver, TP, etc.)	No	Implicit (Safe/Unsafe Context)

<i>Benchmark</i>	<i>Task Focus</i>	<i>Evaluation Method</i>	<i>Verifiability</i>	<i>Context Handling</i>	<i>Norm/Rule Representation</i>	<i>Interference Testing?</i>	<i>Normality Manipulation?</i>
<i>RuleBench</i>	Inferential Rule Following	Accuracy + Error Analysis (Triggering/Execution)	High (vs. Logic)	Explicit Facts/Rules in Prompt	Explicit Inferential Rules (NL / Formal)	Implicit (Noise Rules)	Counterfactual Rules
<i>RuleTaker</i>	Deductive Reasoning (NL Theory)	Classification Accuracy	High (vs. Logic)	Explicit Facts/Rules in Prompt	Explicit Deductive Rules (NL Templates)	No	No
<i>ProofWriter</i>	Proof Generation / Abduction (NL Theory)	Proof Verification / Accuracy	High (Proof Check)	Explicit Facts/Rules in Prompt	Explicit Deductive Rules (NL)	No	No
<i>Framing the Game</i>	Context Framing Effects (Decision Making)	Analysis of Decision Patterns	N/A (Behavioral)	Narrative Vignettes (Varied Framing)	Implicit (Framing influences perceived norms/goals)	Implicit (Framing)	Yes (Framing Variation)
<i>LongBench</i>	Long Context Understanding (Multiple Tasks)	Task-Specific Metrics (Accuracy, ROUGE, F1)	Varies by Task	Long Input Sequences (Documents, Code, etc.)	N/A (Focus on processing, not adherence)	Implicit (Length)	No
<i>FaithEval</i>	Contextual Faithfulness (RAG Scenarios)	Accuracy (vs. Context)	Medium (LLM/Human Validated)	Challenging Contexts (Unanswerable, Inconsistent, Counterfactual)	N/A (Focus on faithfulness to context)	Implicit (Conflict)	Counterfactual Context
<i>ContextualJudge-Bench</i>	Contextual LLM-as-Judge Evaluation	Accuracy (vs. Gold Preference)	Medium (Human/Model Annotated)	Contextual Scenarios (RAG, Summarization)	Explicit Evaluation Criteria (Faithfulness, Refusal, etc.)	No	No
<i>PrivaCI-Bench</i>	Contextual Privacy Compliance	Accuracy (vs. Legal Standards / MCQs)	High (vs. Legal/QA)	CI-Formalized Contexts (Cases, Policies, Synthetic)	Explicit (CI Norms + Legal Regulations)	No	No
<i>SafeDialBench</i>	Safety vs. Jailbreak (Multi-Turn Dialogue)	LLM/Human Scoring (Identify, Handle, Consistency)	Medium (LLM/Human)	Multi-Turn Dialogue History	Explicit (Safety Taxonomy Dimensions)	Implicit (Jailbreak)	No
<i>Constitutional AI</i>	Alignment Method	Internal Self-Critique	N/A	Internal Processing	Explicit (Constitution Principles)	N/A	N/A

Note: ‘Interference Testing?’ and ‘Normality Manipulation?’ columns refer to whether the benchmark explicitly and systematically tests these aspects as a core part of its design.

about social behaviour is crucial. Evaluating performance in these scenarios could reveal important insights into model flexibility, generalization capabilities, potential biases (e.g., rigidly adhering to pre-trained norms despite explicit contextual information), and overall robustness when encountering novel social situations. The current lack of such evaluations limits our understanding of LLM adaptability beyond known normative patterns.

Weaknesses in Measuring Interference Effects

While various benchmarks evaluate adherence to instructions, norms, or rules in isolation or combination (e.g., ComplexBench), there is a notable weakness in methods designed to explicitly measure and quantify *interference effects* when these different types of constraints conflict within a context. Performance degradation is often observed when complexity increases (FollowBench), context lengthens (LongBench), or rules are counterfactual (RuleBench), but the specific contribution of *conflict* between, for example, a direct instruction and an implicit contextual norm, is typically not isolated from other factors like general cognitive load or processing limitations. Benchmarks like FaithEval implicitly test interference by presenting inconsistent or counterfactual contexts, but don't necessarily quantify the interference effect itself.

The central research question regarding methods to measure interference or conflicts between context-derived norms and standard instruction-following remains largely unanswered by existing benchmarks. There is a need for objective and verifiable methods to quantify how much adherence to an instruction degrades specifically due to the introduction of a conflicting contextual norm or rule. For example, if an LLM follows instruction 'A' with 90% accuracy in a neutral context, what is its accuracy on 'A' when introduced into a context with norm 'B' that implicitly discourages action 'A'? Quantifying such interference effects is vital for understanding the trade-offs inherent in LLM alignment — such as balancing helpfulness (following instructions) with harmlessness or appropriateness (adhering to norms) — and for developing models that can gracefully negotiate conflicting demands.

Insufficient Isolation and Explanation of Performance Degradation

Related to the lack of interference measurement, there is insufficient isolation and explanation of *why* performance degrades in context-rich scenarios where norms and instructions might conflict. While benchmarks show that performance drops occur [273], they often lack the fine-grained analysis needed to attribute the failure to specific causes in conflict situations. RuleBench's distinction between Triggering and Execution errors for rule-following is a valuable step in this direction [253], but similar breakdowns are needed for norm-instruction conflicts.

When an LLM fails to follow an instruction in a normatively complex context, the reason could be multifaceted: did it fail to recognize the relevant norm? Did it prioritize the norm over the instruction (or vice-versa)? Was it unable to find a compliant resolution satisfying both? Or did the combined complexity simply exceed its processing capabilities? Current

evaluation methodologies generally do not provide mechanisms to disentangle these potential failure modes. Without a clearer understanding of *why* adherence breaks down in specific conflict scenarios, developing targeted interventions — whether through improved prompting strategies, specialized fine-tuning, or architectural modifications — remains challenging.

6.10 Opportunities for Future Research

The methodological gaps identified in the current landscape of LLM evaluation present clear opportunities for future research aimed at developing more robust, context-aware, and reliable models.

Benchmark Design for Isolating Interference

A promising direction is the development of novel benchmark designs specifically engineered to isolate and measure interference effects between contextual norms and explicit instructions. This requires moving beyond evaluating adherence to each constraint type in isolation. Future benchmarks could draw inspiration from the counterfactual setups in RuleBench or the context manipulation in Casa and CASE-Bench, but with a specific focus on creating controlled conflicts. For instance, researchers could design paired scenarios: one with a neutral context and an instruction, and another with the same instruction but embedded in a narrative context where an explicitly defined (potentially unusual) norm conflicts with the instruction. By carefully matching the complexity of the instruction and the length/complexity of the context across paired scenarios, the performance difference could be more reliably attributed to the norm-instruction conflict. Developing datasets with such controlled pairings would enable systematic quantification of interference effects and the trade-offs involved in navigating conflicting demands.

Objective Metrics for Contextual Norm Adherence

Improving the objectivity and verifiability of contextual norm adherence evaluation is another critical area. While norms are often subjective, certain aspects might be amenable to more objective measurement. Future work could explore adapting verifiable evaluation techniques from related domains. For example, methods from IFEval (script-based checks for verifiable aspects like avoiding forbidden topics or using specific address terms dictated by context) or ProofWriter/RuleBench (verifying logical consistency with stated contextual rules or norms) could be applied where appropriate. Hybrid approaches, combining objective checks for verifiable components with nuanced LLM-as-judge frameworks (like RAL) or structured human evaluation for more subjective aspects, warrant further investigation. Additionally, exploring metrics beyond simple accuracy, such as the model’s expressed confidence calibration when making norm-related judgments or its consistency in applying norms

across semantically similar but superficially different contexts, could provide deeper insights into the reliability of its normative reasoning.

Systematic Study of Abnormal/Counter-Intuitive Norms

A largely untapped area for research involves systematically studying LLM behaviour when confronted with contextual norms that are explicitly defined as abnormal or counter-intuitive within a specific narrative setting. This extends the idea of counterfactual rules (RuleBench) and counterfactual contexts (FaithEval) to the realm of social norms. Research could investigate how LLMs respond when a story explicitly states that, *in this specific context*, a typically rude behaviour is considered polite, or a normally acceptable action is forbidden. Factors influencing adherence in such scenarios could be systematically varied, including the explicitness and perceived strength of the abnormal norm, the degree of conflict with typical expectations, and the framing of the narrative (drawing from Framing the Game). Analysing LLM responses in these situations could reveal crucial information about their flexibility, their reliance on pre-trained statistical norms versus explicit contextual information, and potential biases or fragilities that only surface when dealing with unexpected social rules. This line of inquiry is essential for understanding how LLMs might behave in genuinely novel or culturally diverse situations not well-represented in their training data.

6.11 Conclusion

The evaluation of large language models' adherence to instructions, norms, and rules within context is a rapidly evolving field, critical for ensuring the safe, reliable, and ethical deployment of these powerful technologies. Significant progress has been made in developing a diverse array of benchmarks, moving from early evaluations of simple instruction following (e.g., IFEval) and logical deduction (e.g., RuleTaker) towards assessing more complex phenomena. These include handling composite instructions (ComplexBench), judging cultural appropriateness in narratives (NormAd), detecting norm violations in dialogues (NormDial), navigating norms in agentic web tasks (Casa), applying inferential rules under varying conditions (RuleBench), generating verifiable proofs (ProofWriter), understanding long contexts (LongBench), evaluating faithfulness to challenging contexts (FaithEval), assessing contextual judgments (ContextualJudgeBench), and making context-dependent safety and privacy judgments based on formal theories like Contextual Integrity (CASE-Bench, PrivaCI-Bench).

Despite this progress, substantial methodological challenges remain. A key tension persists between the need for objective, verifiable evaluation (often limiting assessment to simpler, verifiable constraints) and the desire to assess adherence in complex, nuanced, real-world scenarios (often requiring less objective LLM or human judgment). Furthermore, critical gaps exist in current methodologies. There is a lack of benchmarks specifically designed to isolate and measure interference effects when instructions, norms, and rules conflict within a context. Methods for systematically manipulating the perceived normality of contextual

norms to probe model adaptability and robustness are underdeveloped. Finally, while performance degradation in complex scenarios is observed, fine-grained analysis to pinpoint the specific causes of failure, particularly in conflict situations, is often missing.

For the dissertation’s broader agenda, the stakes of these evaluation gaps extend well beyond ‘general alignment’. Privacy, too, is a contextual performance: whether disclosing a phone number, forwarding an e-mail, or summarising a private medical record is *appropriate* hinges on who is asking, for what purpose, and under which social or legal constraints. Chapter 4 showed how lineage-scrubbing at collection time dissolves those contextual constraints; Chapter 5 revealed that people decide whether to invoke ‘forget-me’ rights only after weighing the *social stakes* for their peers and the institutional setting of the model, underscoring that erasure is itself a contextual, relational judgement rather than a purely technical toggle. Chapter 6 now exposes the same fracture at *utilisation*: current adherence benchmarks rarely test whether an LLM can **recognise and obey privacy norms that emerge from context**. When a model answers a ‘harmless’ trivia prompt by leaking another user’s e-mail, the failure is not one of raw accuracy but of contextual privacy judgement. Yet none of the instruction or rule suites, and only a minority of the norm-or safety-focused suites (e.g. CASE-Bench, PrivaCI-Bench), encode the Contextual-Integrity parameters that would let us detect such leaks. Until evaluation protocols track *who may transmit what to whom*, privacy guarantees will remain as brittle at deployment as they are at collection and erasure. In that sense, the contextual-evaluation gap completes the dissertation’s diagnosis: individual-control tools fracture first at data provenance, then at post-hoc deletion, and finally at real-time interaction — precisely where privacy expectations are most situational.

Addressing these gaps is crucial for advancing the field. Future research should focus on developing sophisticated benchmark designs that allow for the isolation of interference effects, exploring objective and hybrid metrics for evaluating contextual norm adherence, and systematically studying LLM behaviour when faced with abnormal or counter-intuitive contextual norms. By pursuing these directions, the research community can develop a deeper understanding of how LLMs navigate the complexities of context-dependent instructions and norms, paving the way for the creation of more robustly aligned, contextually aware, and ultimately more beneficial AI systems.

Chapter 7

Discussion

7.1 Restating the Core Tension

Privacy law in the United States still pivots on what Solove terms the *privacy-self-management bargain*: if an organisation furnishes notice, secures consent, and offers a post-hoc opt-out, then further data processing is presumptively legitimate and socially acceptable [247]. Technical research has evolved in lock-step with that legal frame. A first wave of ‘privacy-by-design’ tools seeks to eliminate the data subject altogether: synthetic-data generators swap each record for a statistically similar imposter [152], while differential-privacy mechanisms bound how much any single row can sway an aggregate statistic [83]. A second wave promises to *retroactively* perfect the bargain. Machine unlearning purports to scrub a subject’s imprint from a trained model whenever consent is revoked [37]; watermarking and model deletion orders give regulators a lever to trace or destroy models tainted by illicit data [141]. Most recently, alignment pipelines fine-tune large language models (LLMs) so that a user’s natural-language instruction functions as a micro-contract that the model must obey [292, 17]. Taken together, these artefacts project a seductive storyline: *perfect the per-person control lever and system-level harms will dissipate*.

Mulligan, Koopman, and Doty’s “Design Wars” study shows that privacy-by-design programmes often succeed at the *single* value they target, but only by off-loading frictions onto the rest of the sociotechnical stack — thereby eroding other, equally important values such as non-discrimination, security, or due-process accountability [188, 187]. They call this the *one-value-at-a-time pathology*: each engineering sprint is framed as a battle won for its focal value, while the collateral damage is deferred to a later battle — one that may never be fought.

The three remedies studied in this dissertation exhibit precisely that pattern. Differential privacy, synthetic substitution, and machine unlearning *do* improve an individuated metric of exposure or autonomy, but in doing so they can actively *obstruct* work on collective harms by (i) erasing provenance that regulators need, (ii) incentivizing designers to treat relational bias as an external cost, and (iii) locking governance debates into per-row accounting metaphors.

The chapters that precede therefore treat each artefact not only as *insufficient* for collective protection but — echoing Design Wars — as a potential antagonist to it when adopted as a stand-alone fix.

The three empirical chapters complicate that storyline.

- **Chapter 4** shows how row-level anonymity can coexist with — or even *magnify* — representational injury. When developers pipe Euro-centric face scans into a synthetic generator, the resulting ‘private’ dataset still skews lighter and younger than global reality, enabling the same downstream mis-recognition that sparked the privacy fix in the first place. The FTC’s model deletion remedy likewise falters once lineage is blurred; if a synthetic face cannot be traced back to the original, deception-based enforcement loses its evidentiary hook [51].
- **Chapter 5** reveals that participants treat machine unlearning not as an isolated act of self-protection but as a *relational* tactic. Interviewees weighed whether unlearning their rent-payment history might amplify unfairness for peers who remained in the training pool, or whether purging medical data would degrade diagnostic performance for demographically similar patients. Such reasoning directly echoes Viljoen’s claim that data governance must grapple with horizontal, population-level externalities rather than vertical, data-subject rights alone [263].
- **Chapter 6** demonstrates that instruction-following scores collapse the moment the same prompt is reframed inside a narrative context that embeds conflicting social norms. A model may flawlessly obey “Place the cup on the table” in isolation yet reject the action when the setting is a funeral, or obey it but add an apology when the cup belongs to someone of higher status. Benchmarks that ignore such framing effects systematically over-estimate alignment — an insight that dovetails with Weidinger et al.’s call for ‘ground-truthing’ evaluation science against lived social contexts [272].

Across these three case-studies a single conclusion crystallises: *artefacts optimised for atomised consent cannot, by construction, perceive the relational harms they help instantiate*. Shelby et al.’s typology reminds us that interpersonal privacy violations are only one of five harm categories — sitting beside representational, allocative, service-quality, and systemic harms [237]. Yet today’s individual-control tools are evaluated almost exclusively on that single axis. The result is a persistent ‘law-tech mismatch’: regulators armed with notice-and-consent doctrines applaud synthetic anonymisation or post-hoc unlearning as compliance milestones, while allocative inequities and cultural erasures proceed unchecked.

The remainder of this Discussion is organised in three steps. Section 7.2 distils the empirical material into four recurrent *failure patterns*. Section 7.3 then asks what happens when those patterns collide with existing enforcement and standards pathways, and sketches measurement-oriented hooks — lineage tags, erasability profiles, context-rich stress tests — that could realign technical levers with regulatory teeth. Finally, Section 7.4 projects the open questions that follow: how a comparative science of socio-technical evaluation

might mature, and where design or policy experimentation could move beyond the limits of individual control.

7.2 Cross-Chapter Synthesis of Failure Modes

Although Chapters 4 – 6 investigate ostensibly different artefacts — *synthetic data pipelines*, *machine-unlearning interfaces*, and *instruction-tuned LLMs* — the empirical results converge on four characteristic ways in which an individual-control lens obscures, rather than resolves, sociotechnical harm. Below we trace each failure mode across the three domains, weaving the chapter-specific findings into a single diagnostic narrative.

Relational Blind Spots

A through-line across the three empirical chapters is that tools calibrated on *row-level privacy or autonomy* routinely overlook harms that arise *between* people who share relevant traits, resources, or social positions.

Synthetic data (Chapter 4).

What our first case study lays bare is not a numeric skew we ourselves quantified, but the *design latitude* wielded by those who build synthetic pipelines. Because the GAN faces in our evaluation were drawn from a generator trained and tuned by the vendor, decisions about which age-ranges, skin-tones, and facial features to preserve, perturb, or invent sat entirely with the dataset curator — not with the data subjects the faces purportedly ‘protect’. The resulting corpus *looked* visually diverse at a glance, yet close inspection revealed that darker-skinned female phenotypes appeared mainly as slight variations on a handful of base meshes, whereas lighter-skinned male phenotypes enjoyed far richer variety. Rather than an accidental sampling artefact, that imbalance illustrates the structural power asymmetry the chapter interrogates: once privacy is defined as per-row de-identification, the generator’s authors are free to decide which demographic cues matter, thereby recoding the original representational gap into the synthetic twin. The argument resonates with Guo et al.’s finding that many off-the-shelf face-generators default to lighter, younger, Euro-centric features even when fed a balanced source set [125].

Machine unlearning (Chapter 5).

Interview participants approached unlearning as a collective — not merely personal — decision. Several initially celebrated the right to erase their rental-payment history, but reconsidered when they realised that large-scale departures could distort the risk pool used to set premiums for neighbours of similar socio-economic status. Such reflections confirm Viljoen’s relational governance thesis: data flows generate obligations across peers and cohorts, not just between a collector and an individual [263]. Formal RTBF doctrine, by

contrast, frames erasure as a purely personal entitlement and offers no language for negotiating those horizontal effects.

Context-sensitive LLM evaluation (Chapter 6).

Large-language models reproduce the same blind spot. On syntactic compliance tests — token counts, banned-word filters — GPT-4 achieved near-perfect scores; yet its social-appropriateness accuracy dropped markedly on stories situated in Ghana, Lebanon, or Bolivia relative to U.S. vignettes. The model satisfied atomistic constraints but missed the relationally situated norm, mirroring the ‘representational harm’ taxonomy introduced by Buolamwini and Gebru [44]. Evaluations that declare success once a single instruction is followed therefore risk masking cultural mis-alignment that unfolds at the narrative level.

Across data generation, post-hoc erasure, and model behaviour, individual-control tools excel at insulating single records yet remain blind to the relational texture of harm. The three chapters supply case material showing how that fallacy persists even when cutting-edge privacy and alignment techniques are applied in good faith.

Provenance Opacity

Synthetic data (Chapter 4).

The synthetic-data pipeline can sever the link between source and output. Once the CASIA-WEBFACE photographs are passed through Synthesis AI’s volumetric mesh and texturing engine, the resulting JPEGs no longer retain a hash, URL, or licence field that would allow observers to trace them back to the scraped celebrity images. In the chapter’s regulatory analysis (4.6) we note that this rupture frustrates “affected work-product” remedies such as the model deletion orders issued in *Everalbum* and *Clearview*: if an artefact cannot be shown to derive from unlawfully collected data, it falls outside the enforcement scope even when it continues to encode the same representational bias.

Machine unlearning (Chapter 5).

Machine-unlearning tool-chains promise to erase a datum’s ‘residual influence’, yet they do so by overwriting gradient contributions without leaving a verifiable audit trail. The chapter’s scenario-elicitation study surfaced this opacity as a practical worry for participants: several asked how an external watchdog could confirm that their e-mail corpus was truly forgotten. The absence of durable lineage artefacts thus shifts the evidentiary burden from the data controller, who alone possesses internal training logs, onto individuals who must prove a negative — that hidden influence still persists.

Context-sensitive LLM evaluation (Chapter 6).

Large-language models amplify the same problem at inference time. As an example, LongBench-v2 shows runs where only 28% of citations emitted by a 13-billion-parameter RAG model actually pointed to the shard that informed its answer; the remainder were hallucinated or linked to dead URLs. With the provenance chain already broken inside the model, downstream actors cannot reconstruct which documents shaped an unsafe or biased response, nor can data subjects invoke erasure rights against the true source.

Together, the three cases show how individual-control fixes gain their apparent efficacy precisely by *rendering provenance unverifiable*, thereby eroding the accountability substrate on which collective safeguards rely.

Contextual Brittleness

Synthetic data (Chapter 4).

Our analysis of Microsoft *Face Synthetics* illustrates how a dataset can look statistically diverse yet remain context-poor. All 100,000 images are permutations of just 511 volumetric head-scans — fewer than thirty labelled Black men and even fewer women from Black, South-Asian, or Middle-Eastern groups. Hair texture, religious head-coverings, low-light noise, and off-angle poses are absent because they cannot be generated from the limited base meshes. The result is a corpus that ‘checks the box’ on numerical parity while omitting the very visual contexts that trigger real-world failures, thereby baking brittleness into any model evaluated or trained on it.

Machine unlearning (Chapter 5).

The interview study underscored context’s power to reshape privacy reasoning even when the *factual prompt remained constant*. When participants encountered the statement “*You don’t know what data the system was trained on*” in the low-stakes *food-delivery* scenario, most treated it as a tolerable ambiguity: “It’s annoying,” one person shrugged, “but worst case I get a bad pizza suggestion.” Presented with the same unknown in an *insurance-pricing* scenario, the mood flipped. Several participants immediately reached for the most extreme remedy — machine unlearning — because, as one put it, “this could hike my neighbour’s premium if the training data is off.” A parallel reversal surfaced in the *landlord* vignette: the identical gap in model transparency provoked worries about rental discrimination and spurred calls for outright deletion rather than a simple opt-out. These shifts show that participants do not carry a single, portable privacy preference; instead, they interpret the very same informational deficit through the relational stakes of each setting — precisely the form of contextual brittleness individual-control tools fail to register.

Context-sensitive LLM evaluation (Chapter 6).

Benchmarks reviewed in the literature survey document the same fragility at the language-model layer. In the FRAMING-THE-GAME dataset, GPT-4 favours cooperation when a payoff matrix is wrapped in a playground story but shifts toward defection when the identical incentives are recast as corporate espionage. LongBench further shows that large models often ignore instructions embedded in the middle of very long contexts, focusing only on the first and last few thousand tokens. Because most industrial scorecards grade keyword obedience or toxicity in isolation, they miss these context-dependent failures.

Across the three artefacts, then, contextual drift — not model size or privacy budget — proves the dominant failure trigger: synthetic faces lose cultural signifiers, unlearning choices mutate under social reframing, and LLMs oscillate with narrative setting or prompt length. Any evaluation that ignores those contextual axes is, at best, an optimistic lower bound on real-world safety.

Enforcement Friction

Synthetic data (Chapter 4).

Finally, the operational overhead introduced by sophisticated individual-control levers risks outstripping the capacity of regulators and affected individuals. The FTC’s *Rite Aid* and *Clearview* cases illustrate the point. Both resulted in model deletion mandates, but only after multi-year forensic negotiations over what exactly counted as an ‘affected’ artefact — negotiations complicated further when defendants argued that intermediate synthetic datasets were shielded as fair-use transformations [141]. Elder & Klein note that such model deletion orders are already rare because they “raise the stakes of litigation beyond what typical FTC budgets can sustain” [85].

Machine unlearning (Chapter 5).

The qualitative study made clear that Right-to-Be-Forgotten remedies are only as credible as the evidence that a model has, in fact, ‘forgotten’. Participants were quick to point out that, once their data had shaped internal weights, they had no practical way to verify removal: “deleting a row is easy,” one interviewee noted, “but how do I know the gradients are gone?” That doubt mirrors a legal vacuum. Current RTBF statutes assume deletion is binary and inspectable, yet machine-unlearning research acknowledges residual influence, architecture-specific feasibility, and probabilistic guarantees. Because there is no agreed-upon audit metric for what counts as ‘erased’, data-protection authorities cannot readily certify compliance or levy fines, and participants in our study voiced little confidence that regulators could back their rights with enforceable tests. The gap between doctrinal expectations and technical verifiability thus becomes an enforcement friction: requests for erasure multiply, but neither subjects nor supervisors can prove the job was done.

Context-sensitive LLM evaluation (Chapter 6).

Benchmark culture introduces a parallel bottleneck for oversight of large language models. Firms now publish alignment scorecards built on context-free metrics — keyword obedience, aggregate toxicity rates, or single-turn refusal counts — and standards bodies such as ISO/IEC 42001 and the NIST AI RMF draft guidance that cites those very numbers. Yet Chapter 6 shows that the same models falter when norms conflict, narratives shift, or long contexts dilute attention — scenarios absent from the self-reported dashboards. Independent auditors struggle to probe these weak spots because core artefacts (system prompts, constitutional rule sets, safety filters) are withheld as trade secrets, and existing public benchmarks do not encode the situational parameters regulators care about. The result is an accountability log-jam: companies can claim compliance by passing shallow tests, while agencies lack both the artefacts and the context-rich evaluations needed to contest those claims.

Take-Away

Each expanded failure mode deepens the empirical case that an individual-centric conception of privacy and alignment not only *misses* key forms of harm but can generate the very kind of collateral damage that Design Wars warns about — shoring up one value while degrading the conditions needed to secure the rest. The following sections of the Discussion therefore pivot from diagnostics to design and policy strategies capable of accounting for these relational, contextual, and operational realities.

7.3 Aligning Technical Levers & Accountability Pathways

The dissertation’s core argument is not only that individual-control solutions leave relational harms untouched; it is that *technical interventions are only as effective as the regulatory and standards regimes that make them auditable*. Chapters 4 – 6 each expose a location where that coupling fails: synthetic data pipelines outrun the FTC’s notion of “affected work product”, machine-unlearning promises outrun the binary logic of statutory erasure rights, and context-free evaluation regimes outrun the accountability demands now taking shape in AI safety standards. This section synthesises those disconnects and sketches concrete hooks — documentation, metrics, and oversight interfaces — that could realign technical levers with enforceable obligations.

Synthetic Pipelines & Model Deletion Orders

Model deletion orders issued under the FTC Act, such as the *Clearview AI* and *Rite Aid* settlements, rest on a seemingly simple premise: if the underlying data were collected unfairly or deceptively, *every artefact* derived from that data — models, embeddings, evaluation

sets — must be deleted [141]. Chapter 4 shows how GAN and diffusion pipelines break the evidentiary chain that makes this remedy work. Once training images are replaced by synthetic ‘twins’, lineage metadata evaporates: the diffusion checkpoint no longer carries a hash that can be matched to any specific celebrity JPEG, and synthetic faces deployed in forensic benchmark appear as novel IP outside BIPA’s reach. Enforcement stalls because regulators cannot *prove* that the derivative artefact is “affected work product”.

A technical fix is already within reach: cryptographic watermarking or provenance tags embedded at generation time could make downstream datasets self-authenticating, restoring the traceability that the FTC’s remedy presupposes [36]. But without a parallel *documentation mandate*, nothing compels firms to keep those tags intact. Proposed U.S. bills such as the *American Data Privacy and Protection Act* and the EU’s final AI Act drafts both contain “duty to document” clauses for high-risk AI systems; coupling watermarking with such legal record-keeping requirements could convert a best-effort provenance signal into an enforceable audit trail.

Machine Unlearning & Statutory Erasure Rights

GDPR Article 17 and the California Consumer Privacy Act (CCPA §1798.105) both give individuals a right to have their personal data ‘erased’ or ‘deleted’. Regulatory guidance still frames that obligation in *binary* terms: a controller must be able to show that the data ‘are no longer present in its systems’ unless a statutory exception applies [91, 46]. Machine-learning artefacts up-end that binary logic: deleting the row does nothing if the parameters learned from it remain intact. Data-protection authorities have not yet defined how much residual signal is tolerable, leaving them without a quantitative hook for fines or corrective orders. Computer-science work has started to fill the gap — e.g. influence-function tests, Shapley Leave-One-Out scores, or the FORGETTING L2 metric surveyed by Lynch et al. [176] — but the thresholds are still ad-hoc and have not (yet) been adopted in enforcement practice.

Evaluation science can play a bridging role here. A harmonised *erasability profile* — analogous to an energy-efficiency label — could specify (i) attack models, (ii) confidence intervals, and (iii) acceptable leakage budgets for different risk classes (medical, financial, public-facing chatbots). Regulators would then be able to cite a “fails grade C on erasability” finding as concrete evidence, turning what is now a probabilistic promise into a justiciable compliance target. Pilot work towards such labels is already underway in the ISO/IEC 42001 AI management system draft; Chapter 5’s metrics provide empirical grounding for that standard-setting process.

Contextual Evaluation & Emerging Safety Standards

Chapter 6 shows that many vendors now present high scores on *context-free* leader-boards — e.g. keyword-based instruction suites such as IFEVAL or single-turn toxicity filters — as de facto evidence of alignment. Yet the same chapter’s survey makes clear that those metrics rarely survive richer scenarios. Benchmarks that extend the prompt into a long

document (LONGBENCH), embed instructions in culturally specific stories (NORMAD), or ask an agent to navigate a web task with conflicting cues (CASA) all report sharp accuracy drops. Because today’s conformity-assessment templates (e.g. ISO/IEC 42001, NIST AI RMF profiles) still reference these context-thin scorecards, the standards pipeline risks certifying models that have never been tested for the very collisions — between length, culture, and competing obligations — that drive real-world harm. In short, the evaluation layer mirrors the individual-control focus of notice-and-consent law: it can verify atomic obedience, but it cannot tell whether a system can negotiate the multi-constraint environments in which privacy, safety, and fairness norms actually operate.

The disconnect is worrisome because standards bodies are beginning to canonise those very metrics. NIST’s *AI Risk Management Framework* and the EU AI Act’s systemic-risk tier both call for *context-specific* assurance, yet their technical annexes still reference instruction-accuracy and toxicity rates drawn from context-thin suites. Drafts of ISO/IEC 42001 and IEEE P7009 likewise cite industry scorecards while providing little guidance on multi-constraint evaluation. The result is an accountability pipeline that risks rubber-stamping systems whose alignment has never been tested against the norm conflict, framing effects, or long-context interference that drive real-world harm.

Two upgrades are needed. *First*, evaluation suites must embed the **transmission parameters** — sender, recipient, information type, and purpose — that Contextual-Integrity theory identifies as determinants of appropriateness. CASE-Bench is a useful prototype, but it covers only safety judgements; similar CI-grounded constructs are required for fairness, privacy, and instructional adherence. *Second*, those context-rich benchmarks must be wired directly into certification regimes. We propose a dual-axis stress-test template: couple long-context interference tasks with norm-conflict vignettes and require models to maintain an acceptable risk budget across both. Such tests would translate the principle-based aspirations of NIST, ISO 42001, and IEEE P7009 into concrete quantitative thresholds — aligning technical levers with the multi-factor accountability pathways that regulators are now mandating.

Take-away

Accountability is only as strong as the *measurement substrates* handed to regulators. Linking technical levers to real-world safeguards is therefore not a matter of tacking on one more dashboard metric; it is a **sociotechnical design problem**. *Lineage tags* generated by synthetic-data pipelines must acquire *legal standing* so that model deletion orders can meaningfully track derivative artefacts. Probabilistic *erasability scores* emerging from machine-unlearning research will need regulatory *thresholds* — a data-protection authority must be able to say, in black-letter law, that a residual-influence rate of $x\%$ is (or is not) acceptable. And the *context-rich evaluations* described above must be written directly into certification workflows such as ISO/IEC 42001 or the EU AI Act’s conformity-assessment modules. Until those bridges are built, individual-control tools will continue to ‘solve’ privacy on paper while leaving relational harm — and public accountability — largely untouched.

The science of evaluation — taken up in the next section — thus becomes the hinge that reconnects these technical levers to enforceable safeguards.

7.4 Future Directions: Evaluation, Design, and Policy

Towards a Science of Socio-Technical Evaluation

Versioned and drift-audited benchmarks.

A reproducible science depends on the ability to specify *which* test a model has actually passed. Static leader-boards obscure the fact that data distributions mutate in the wild: face datasets accumulate new lighting artefacts as cameras improve; cultural-norm vignettes age as slang evolves; and long-context corpora undergo URL rot. Without formal versioning, a high score on LongBench may refer to a context length that has since doubled or to documents that no longer exist, rendering cross-paper comparisons meaningless.

Embedding semantic-version tags and obligatory changelogs would let researchers trace when and why a model’s performance shifted, and would give auditors an anchor for time-stamped compliance checks. The practice could borrow directly from software engineering: a minor version increment for typo fixes, a major bump when a demographic split is added or removed. Regulators could then cite ‘LongBench v2.3’ in licensing or procurement documents with the same precision that pharmaceutical agencies cite Phase-III trial protocols. Absent such drift audits, we risk repeating the ImageNet story — years of apparent accuracy gains that turned out to be artefacts of unreported dataset revisions.

Power-aware protocol design.

Most fairness and safety studies still rely on convenience samples of 100–200 prompts, despite effect sizes that require an order of magnitude more to reach conventional statistical power [272]. The mismatch allows vendors to claim ‘no significant difference’ simply because the test could not have detected a real-world gap. A power-aware protocol begins with a pre-registered effect size (e.g. five-point disparity in refusal rate) and then computes the minimum number of prompts, annotators, and model replications needed to detect that gap with 95% confidence.

Publishing these calculations alongside the benchmark would shift the burden of proof. A developer who wishes to rely on a smaller sample must justify a different effect size or accept a wider confidence interval. For regulators, power annotations serve as a decision heuristic: any model submitted for certification using an under-powered test can be flagged for supplementary auditing before deployment. The approach mirrors epidemiology, where ethical review boards reject studies that are too small to yield actionable insight; it merely translates that ethic to the AI domain.

Transparent annotation pipelines.

Human raters remain indispensable for judging contextual norm adherence, but the labour conditions under which they work shape the labels they produce. Low pay, ambiguous guidelines, or culturally homogeneous pools can induce systematic bias — exactly the problem the benchmarks aim to expose. Building on dataset-genealogy work by Scheuerman et al. [228], every evaluation release could include a ‘rater datasheet’: hourly wage, demographic breakdown, training materials, and inter-rater reliability scores.

Such transparency would let downstream researchers quantify how much of a performance delta stems from model behaviour versus annotator composition. It would also facilitate replicability studies that intentionally swap rater pools to probe cultural robustness. Finally, disclosure creates an incentive for benchmark owners to pay living wages, as undercompensation becomes a reputational risk once documented in the public record.

Context-Conflict Evaluation as open research.

Our case studies repeatedly show that models pass atomic tests yet unravel when instructions and norms collide. A promising — but deliberately open-ended — avenue is a *Context-Conflict Evaluation Suite* (CCES). Instead of static prompts, a CCES would generate matched pairs: one scene where an instruction is compliant with the background norm, and a second where the same instruction violates an explicitly stated, possibly *counter-typical*, norm. The difference in success rates forms an *interference index*.

Crucially, the suite would vary *normality* along a gradient: statistically typical, merely unusual, and culturally taboo. That design lets us plot a reconciliation curve — how gracefully a model re-writes, refuses, or asks clarifying questions as the norm-instruction gap widens. Although Chapter 6 sketches early ingredients, developing a CCES will require collaborative work on scenario libraries, adversarial prompt generation, and agreement metrics that remain reliable across cultures. If successful, the index could join existing safety and bias metrics in policy frameworks such as the EU AI Act’s conformity-assessment or NIST’s AI Risk Management Profile.

Design Experiments Inspired by the Findings

Secure provenance ledgers for synthetic workflows.

Chapter 4 showed how GAN pipelines sever legal traceability: once a face is synthesised, neither data subjects nor regulators can demonstrate lineage. Embedding cryptographic hashes of every intermediate artefact — base mesh, texture map, trained generator weights — into an append-only ledger would restore that chain of custody. The ledger could be anchored on a permissioned blockchain or a simpler Merkle-tree structure maintained by an independent registrar, ensuring that any subsequent model inherits a verifiable provenance trail.

From a user-interface standpoint, dataset cards would expose a tamper-proof `root hash`. Auditors could request random challenges: “Reveal the mesh and training commit that hash to this digest.” Failure to comply would indicate an unauthorised substitution. Such infrastructure turns ‘trust me, it’s synthetic’ into a falsifiable claim, lowering the evidentiary burden in model deletion or licensing disputes.

User-facing erasability dashboards.

Participants in Chapter 5 struggled to gauge whether unlearning actually removed their influence from a model. An erasability dashboard would translate complex gradient-retraining statistics into an intuitive dial: e.g. “Residual influence: $3.2\% \pm 0.4\%$ at 95% confidence”. Behind the scenes, the system would run a battery of membership-inference probes on shadow models fine-tuned to the user’s data slice and report the mean leakage.

Making that metric public accomplishes two goals. First, it gives data subjects a basis for informed consent when weighing deletion versus unlearning. Second, it furnishes regulators with a quantitative threshold akin to differential-privacy ϵ . A supervisory authority could then mandate an influence ceiling — say, no single record may contribute more than 0.1% of logits — creating a concrete compliance target instead of the current binary fiction that data can be ‘fully erased’.

Collective-control hooks in unlearning APIs.

The right-to-be-forgotten doctrine remains individualistic, yet our study participants often asked, “What happens to my neighbours if I opt out?” Future unlearning APIs could accept cohort-level requests, allowing a tenants’ union or patient-advocacy group to withdraw data en masse when collective risk outweighs benefit. Technically, this requires a reconciliation algorithm: if members of a cohort submit conflicting commands, the API must honour the most restrictive request or escalate to human review.

Governance-wise, such hooks would institutionalise the relational privacy logic articulated by Viljoen: data is co-produced and its removal should be a community decision. Early prototypes might integrate with *data trusts*, where trustees exercise the withdrawal right on behalf of the group. By shifting the locus of decision-making, the design turns unlearning from a purely personal protective measure into a bargaining chip for collective leverage over data-hungry platforms.

Policy Exploration Road-Map

Lineage-disclosure mandates for synthetic data.

Legislators can codify what the provenance ledger above merely facilitates. A disclosure rule would compel any entity releasing synthetic corpora — commercial or open source — to attach a machine-readable manifest of source licences, transformation steps, and hashing

proofs. Non-compliance would expose the publisher to statutory damages analogous to those under the GDPR for unlawful data processing.

Such a mandate makes synthetic data traceable under existing biometric or copyright regimes rather than inventing new categories. It also aligns with draft U.S. AI bills that already contemplate “documentation of data origin” as a condition for safe-harbour protections. The legal stick thereby reinforces the technical carrot of provenance-ledger tooling.

Regulatory ‘forgettability audits’.

Privacy commissioners currently accept unverified attestations that data has been deleted or unlearned. Borrowing from financial stress-tests, regulators could require annual ‘forgettability audits’ that publish residual-influence curves across model layers and training checkpoints. Audits would be performed by accredited third-party labs, similar to penetration-testing in cybersecurity.

The public reports would let civil-society groups benchmark vendors and advocate for stricter thresholds where necessary. More importantly, they would give courts an objective yardstick when adjudicating whether a firm has met its erasure obligations, replacing today’s adversarial duelling of expert affidavits.

State-funded benchmark registries.

Just as clinical-trial registries prevent p-hacking by locking protocols before data collection, a national or international registry for AI benchmarks would host versioned datasets, evaluation scripts, and power analyses. Researchers would pre-register which task split and metric they plan to report, reducing post-hoc cherry-picking.

Public funding ensures that benchmarks addressing low-resource languages, disability contexts, or marginalised cultures receive maintenance budgets comparable to marquee leader-boards. By tying grant eligibility to open registration, agencies can nudge the community toward the evaluation best practices outlined in §7.1 without new legislation.

Collective redress for representational harm.

Current legal doctrine funnels most AI litigation through privacy or consumer-fraud statutes, side-lining representational injuries such as stereotyping or exclusion. One policy experiment is to extend class-action standing to such harms, allowing affected groups to seek injunctive relief or damages without having to quantify individual monetary loss.

Courts already entertain ‘stigmatic injury’ in civil-rights cases; adapting that logic to algorithmic contexts would close the enforcement gap highlighted across all three chapters. Complementary administrative routes include empowering civil-rights offices within data-protection authorities to open investigations absent a privacy trigger, ensuring that relational harms have a pathway to accountability even when lineage is intact and individual consent boxes are ticked.

Together, these forward-looking strands — a rigorous science of evaluation, design interventions that operationalise relational accountability, and exploratory policy pilots — outline a research and governance agenda capable of transcending the paper victories of individual-control privacy. They translate the dissertation’s diagnostic work into a programme for action, one that couples technical levers to socio-legal safeguards at scale.

Chapter 8

Conclusion

This dissertation set out to interrogate the limits of the individual-control paradigm as it manifests in contemporary machine-learning practice and regulation. By following three real-world artefacts — *synthetic-data pipelines*, *machine-unlearning interfaces*, and *instruction-tuned language models* — the chapters revealed four recurring blind spots:

1. Relational blind spots: row-level privacy guarantees often mask group-level misrepresentation and allocative harm.
2. Lineage opacity: technical transformations break the evidentiary chain that both researchers and regulators rely on.
3. Contextual brittleness: benchmarks that ignore situational framing or long-context interference overstate real-world robustness.
4. Enforcement friction: the cost and complexity of auditing individual-control tools frequently exceed institutional capacity, leaving harms un-redressed in practice.

Importantly, the aim has been diagnostic rather than prescriptive: to map where existing technical levers and legal doctrines fall short once relational harm and contextual nuance are brought into view. Highlighting such gaps is a prerequisite to any credible programme of reform.

Beyond one-off ‘tech fixes’. None of the chapters suggests that a better algorithmic primitive — stronger differential privacy, faster unlearning, or cleverer prompts — can single-handedly cure systemic injustice. Each case study instead shows that technical tools are inseparable from the institutions, incentives, and power relations in which they operate. Treating privacy as a purely engineering problem obscures whose values guide data generation, whose interests shape consent defaults, and whose losses are rendered invisible by current audit regimes. If the dissertation critiques existing fixes, it is not to dismiss technology, but to resist the idea that a new line of code can substitute for collective negotiation about fairness, accountability, and redress.

What the dissertation does advocate for. First, embed any technical safeguard in participatory governance: lineage-preserving synthetic standards crafted with affected communities, unlearning protocols paired with collective-impact audits, and interference-aware metrics that surface conflicts among users, not just per-prompt accuracy. Second, shift legal doctrine from a narrow focus on notice-and-consent toward fiduciary and relational obligations — duties that recognise data flows as shared social infrastructure rather than privately owned rows. Third, invest in measurement infrastructures that regulators can actually wield: transparent provenance records, benchmark suites that model cross-context spill-overs, and audit logs that external researchers can inspect without prohibitive cost.

A second, cross-cutting contribution is methodological. Each empirical chapter demonstrated that the way we *measure* adherence — whether to privacy guarantees, erasure rights, or safety norms — profoundly shapes what we can hold systems accountable for. The dissertation therefore joins recent calls for a ‘science of socio-technical evaluation’: reproducible, statistically powered, and context-rich test suites that can expose interference effects and atypical scenarios that static leaderboards miss. Rather than propose a definitive framework, the work identifies the empirical ingredients — norm–instruction conflict matrices, lineage audits, multi-party simulations — that future research must assemble.

What, then, should readers take forward? First, that technical privacy mechanisms cannot be evaluated in isolation from the social relations they reconfigure. Second, that regulatory hard-law tools such as model deletion orders or erasure rights presuppose measurement infrastructures that do not yet exist. And third, that bridging those gaps will require sustained, interdisciplinary collaboration among computer scientists, lawyers, designers, and affected communities.

By documenting where contemporary practice falters, the dissertation provides a roadmap for the next phase of inquiry — one in which technical ingenuity and legal imagination are marshalled not toward ever finer individual control, but toward relational, contextual, and collective accountability.

Bibliography

- [1] Martín Abadi et al. “Deep Learning with Differential Privacy”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Oct. 2016, pp. 308–318. DOI: 10.1145/2976749.2978318. arXiv: 1607.00133 [stat].
- [2] Julio Abascal and Colette Nicolle. “Moving towards Inclusive Design Guidelines for Socially and Ethically Aware HCI”. In: *Interacting with Computers* 17.5 (Sept. 2005), pp. 484–505. ISSN: 1873-7951. DOI: 10.1016/j.intcom.2005.03.002.
- [3] Rediet Abebe and Kira Goldner. “Mechanism Design for Social Good”. In: *AI Matters* 4.3 (Oct. 2018), pp. 27–34. ISSN: 2372-3483. DOI: 10.1145/3284751.3284761.
- [4] John M. Abowd and Ian M. Schmutte. “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices”. In: *American Economic Review* 109.1 (Jan. 2019), pp. 171–202. ISSN: 0002-8282. DOI: 10.1257/aer.20170627.
- [5] William Agnew et al. *The Illusion of Artificial Inclusion*. Feb. 2024. DOI: 10.1145/3613904.3642703. arXiv: 2401.08572 [cs].
- [6] Philip E. Agre. “Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI”. In: *Social Science, Technical Systems, and Cooperative Work*. Ed. by Geoffrey C. Bowker, Susan Leigh Star, and William Turner. Erlbaum, 1997, pp. 131–157.
- [7] Sara Ahmed. “A Phenomenology of Whiteness”. In: *Feminist Theory* 8.2 (Aug. 2007), pp. 149–168. ISSN: 1464-7001, 1741-2773. DOI: 10.1177/1464700107078139.
- [8] Ifeoma Ajunwa. “The Paradox of Automation as Anti-bias Intervention”. In: *Cardozo Law Review* 43 (2022), pp. 1141–1183.
- [9] Ali Alkhatib and Michael Bernstein. “Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, May 2019, pp. 1–13. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300760.
- [10] McKane Andrus and Sarah Villeneuve. “Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. New York, NY, USA: Association for Computing Machinery, June 2022, pp. 1709–1721. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533226.

- [11] Blaise Aguera y Arcas. *Do Algorithms Reveal Sexual Orientation or Just Expose Our Stereotypes?* Jan. 2018.
- [12] Matthew Arnold et al. “FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 6–1. DOI: 10.1147/JRD.2019.2942288.
- [13] Sherry R. Arnstein. “A Ladder Of Citizen Participation”. In: *Journal of the American Institute of Planners* 35.4 (July 1969), pp. 216–224. ISSN: 0002-8991. DOI: 10.1080/01944366908977225.
- [14] Samuel A. Assefa et al. “Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls”. In: *Proceedings of the First ACM International Conference on AI in Finance*. ICAIF ’20. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 1–8. ISBN: 978-1-4503-7584-9. DOI: 10.1145/3383455.3422554.
- [15] Andrea Atzori et al. “The Impact of Balancing Real and Synthetic Data on Accuracy and Fairness in Face Recognition”. In: (Sept. 2024). arXiv: 2409.02867 [cs.CV].
- [16] No Author. *Americans Consider Certain Kinds of Data to Be More Sensitive than Others*. Nov. 2014.
- [17] Ching-Yuan Bai et al. “On Training Sample Memorization: Lessons from Benchmarking Generative Modeling with a Large-scale Competition”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD ’21. New York, NY, USA: Association for Computing Machinery, Aug. 2021, pp. 2534–2542. ISBN: 978-1-4503-8332-5. DOI: 10.1145/3447548.3467198.
- [18] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. Dec. 2022. DOI: 10.48550/arXiv.2212.08073. arXiv: 2212.08073 [cs].
- [19] Kenneth A. Bamberger and Deirdre K. Mulligan. *Privacy on the Ground: Driving Corporate Behavior in the United States and Europe*. Cambridge, MA: MIT Press, 2015.
- [20] Louise Barkhuus. “The Mismeasurement of Privacy: Using Contextual Integrity to Reconsider Privacy in HCI”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Austin Texas USA: ACM, May 2012, pp. 367–376. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207727.
- [21] Julia Barnett and Nicholas Diakopoulos. “Crowdsourcing Impacts: Exploring the Utility of Crowds for Anticipating Societal Impacts of Algorithmic Decision Making”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. July 2022, pp. 56–67. DOI: 10.1145/3514094.3534145. arXiv: 2207.09525 [cs].
- [22] Solon Barocas and Helen Nissenbaum. “Big Data’s End Run around Procedural Privacy Protections”. In: *Communications of the ACM* 57.11 (Oct. 2014), pp. 31–33. ISSN: 0001-0782. DOI: 10.1145/2668897.

- [23] Solon Barocas and Andrew D. Selbst. “Big Data’s Disparate Impact”. In: *California Law Review* 104.3 (2016), pp. 671–732. ISSN: 0008-1221. JSTOR: 24758720.
- [24] Lemi Baruh and Mihaela Popescu. “Big Data Analytics and the Limits of Privacy Self-management”. In: *New Media & Society* 19.4 (2017), pp. 585–601. DOI: 10.1177/1461444815614001.
- [25] Shubhajit Basak et al. “Learning 3D Head Pose From Synthetic Data: A Semi-Supervised Approach”. In: *IEEE Access* 9 (2021), pp. 37557–37573. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3063884.
- [26] Emily Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *FAccT* (2021). DOI: 10.1145/3442188.3445922.
- [27] Ruha Benjamin. *Race after Technology*. Polity Press, 2019.
- [28] Ruha Benjamin. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity, 2019. ISBN: 978-1-5095-2643-7.
- [29] Sebastian Benthall and Os Keyes. “Racial Categories in Machine Learning”. In: *FAccT 2019*. 2019, pp. 289–298. DOI: 10.1145/3287560.3287575.
- [30] Glen Berman. “Machine Learning Practices and Infrastructures”. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 466–481. ISBN: 979-8-4007-0231-0. DOI: 10.1145/3600211.3604689.
- [31] Abeba Birhane and Vinay Prabhu. “Large Image Datasets: A Pyrrhic Win for Computer Vision?” In: *ICCV Workshop on Responsible Vision*. 2021. DOI: 10.1109/WACV48630.2021.00158.
- [32] Abeba Birhane and Vinay Uday Prabhu. “Large Image Datasets: A Pyrrhic Win for Computer Vision?” In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, Jan. 2021, pp. 1536–1546. ISBN: 978-1-6654-0477-8. DOI: 10.1109/WACV48630.2021.00158.
- [33] Abeba Birhane et al. “Power to the People? Opportunities and Challenges for Participatory AI”. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. Oct. 2022, pp. 1–8. DOI: 10.1145/3551624.3555290. arXiv: 2209.07572 [cs].
- [34] Rena Bivens. “The Gender Binary Will Not Be Deprogrammed: Ten Years of Coding Gender on Facebook”. In: *New Media & Society* 19.6 (June 2017), pp. 880–898. ISSN: 1461-4448. DOI: 10.1177/1461444815621527.
- [35] Iffat Borhan and Akhilesh Bajaj. “The Effect of Prompt Types on Text Summarization Performance with Large Language Models”. In: *Journal of Database Management* 35.1 (2024), pp. 1–23. DOI: 10.4018/JDM.358475.
- [36] Wassim Bouaziz, El-Mahdi El-Mhamdi, and Nicolas Usunier. “Data Taggants: Dataset Ownership Verification via Harmless Targeted Data Poisoning”. In: *arXiv preprint arXiv:2410.09101* (2024). arXiv: 2410.09101.

- [37] Lucas Bourtoutle et al. “Machine Unlearning”. In: *2021 IEEE Symposium on Security and Privacy (SP)*. May 2021, pp. 141–159. DOI: 10.1109/SP40001.2021.00019.
- [38] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things out: Classification and Its Consequences*. Inside Technology. Cambridge, Mass: MIT Press, 1999. ISBN: 978-0-262-02461-7.
- [39] Samuel R. Bowman and George E. Dahl. *What Will It Take to Fix Benchmarking in Natural Language Understanding?* Oct. 2021. DOI: 10.48550/arXiv.2104.02145. arXiv: 2104.02145 [cs].
- [40] danah boyd and Kate Crawford. “Critical Questions for Big Data”. In: *Information, Communication & Society* 15.5 (2012), pp. 662–679. DOI: 10.1080/1369118X.2012.678878.
- [41] Tone Bratteteig and Ina Wagner. “Unpacking the Notion of Participation in Participatory Design”. In: *Computer Supported Cooperative Work* 25.6 (Dec. 2016), pp. 425–475. ISSN: 0925-9724. DOI: 10.1007/s10606-016-9259-4.
- [42] Hannah Brown et al. “What Does It Mean for a Language Model to Preserve Privacy?” In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM, June 2022, pp. 2280–2292. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3534642.
- [43] Justin B. Bullock. “Artificial Intelligence, Discretion, and Bureaucracy”. In: *The American Review of Public Administration* 49.7 (Oct. 2019), pp. 751–761. ISSN: 0275-0740. DOI: 10.1177/0275074019856123.
- [44] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, Jan. 2018, pp. 77–91.
- [45] William Cai et al. “Adaptive Sampling Strategies to Construct Equitable Training Datasets”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. New York, NY, USA: Association for Computing Machinery, June 2022, pp. 1467–1478. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533203.
- [46] California Department of Justice. *Final Text of Regulations: California Consumer Privacy Act of 2018*. 2020.
- [47] California Legislature. *California Consumer Privacy Act*. 2018.
- [48] M. Calo. “The Boundaries of Privacy Harm”. In: *86 Indiana Law Journal* 1131 (2011) 86.3 (July 2011). ISSN: 0019-6665.
- [49] Hanqun Cao et al. *A Survey on Generative Diffusion Model*. Dec. 2023. arXiv: 2209.02646 [cs].

- [50] Hongye Cao et al. *SafeDialBench: A Fine-Grained Safety Benchmark for Large Language Models in Multi-Turn Dialogues with Diverse Jailbreak Attacks*. Feb. 2025. DOI: 10.48550/arXiv.2502.11090. arXiv: 2502.11090 [cs].
- [51] Nicholas Carlini et al. *Extracting Training Data from Diffusion Models*. Jan. 2023. DOI: 10.48550/arXiv.2301.13188. arXiv: 2301.13188 [cs].
- [52] M. A. P. Chamikara et al. “Privacy Preserving Face Recognition Utilizing Differential Privacy”. In: *Computers & Security* 97 (Oct. 2020), p. 101951. ISSN: 0167-4048. DOI: 10.1016/j.cose.2020.101951.
- [53] Janet X. Chen, Francesco Vitale, and Joanna McGrenere. “What Happens After Death? Using a Design Workbook to Understand User Expectations for Preparing Their Data”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, May 2021, pp. 1–13. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445359.
- [54] Min Chen et al. “When Machine Unlearning Jeopardizes Privacy”. In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 896–911. ISBN: 978-1-4503-8454-4. DOI: 10.1145/3460120.3484756.
- [55] Yu Ying Chiu et al. *CulturalBench: A Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs*. Oct. 2024.
- [56] Rohit Chopra and Samuel A. A. Levine. *The Case for Resurrecting the FTC Act’s Penalty Offense Authority*. SSRN Scholarly Paper. Rochester, NY, Oct. 2020. DOI: 10.2139/ssrn.3721256. Social Science Research Network: 3721256.
- [57] Paul Christiano et al. *Deep Reinforcement Learning from Human Preferences*. Feb. 2023. DOI: 10.48550/arXiv.1706.03741. arXiv: 1706.03741 [cs, stat].
- [58] Angèle Christin. “Algorithms in Practice: Comparing Web Journalism and Criminal Justice”. In: *Big Data & Society* 4.2 (Dec. 2017), p. 2053951717718855. ISSN: 2053-9517. DOI: 10.1177/2053951717718855.
- [59] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. Dec. 2022. DOI: 10.48550/arXiv.2210.11416. arXiv: 2210.11416 [cs].
- [60] Danielle Citron and Daniel J. Solove. “Privacy Harms”. In: *Boston University Law Review* 102 (2021), pp. 793–858.
- [61] Aloni Cohen and Kobbi Nissim. “Towards Formalizing the Gdpr’s Notion of Singling Out”. In: *Proceedings of the National Academy of Sciences* 117.15 (2020), pp. 8344–8352. DOI: 10.1073/pnas.1914598117.
- [62] Aloni Cohen et al. *Control, Confidentiality, and the Right to Be Forgotten*. Oct. 2022. DOI: 10.48550/arXiv.2210.07876. arXiv: 2210.07876 [cs].
- [63] Julie E. Cohen. *Between Truth and Power: The Legal Construction of Informational Capitalism*. New York: Oxford University Press, 2019.

- [64] Commission Nationale de l'Informatique et des Libertés (CNIL). *Guide d'audit Des Modèles d'intelligence Artificielle*. Nov. 2023.
- [65] Committee on Privacy and Kenneth Younger. *Report of the Committee on Privacy*. Tech. rep. Cmnd. 5012. London: Her Majesty's Stationery Office, 1972.
- [66] Eric Corbett, Emily Denton, and Sheena Erete. "Power and Public Participation in AI". In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–13. ISBN: 979-8-4007-0381-2. DOI: 10.1145/3617694.3623228.
- [67] Nick Couldry and Ulises A. Mejias. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press, 2019.
- [68] Lorrie Faith Cranor, Joseph Reagle, and Mark S. Ackerman. *Beyond Concern: Understanding Net Users' Attitudes About Online Privacy*. Apr. 1999. DOI: 10.48550/arXiv.cs/9904010. arXiv: cs/9904010.
- [69] Kate Crawford and Trevor Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets". In: *AI & SOCIETY* 36.4 (Dec. 2021), pp. 1105–1116. ISSN: 1435-5655. DOI: 10.1007/s00146-021-01162-8.
- [70] Zihang Dai et al. *Good Semi-supervised Learning That Requires a Bad GAN*. Nov. 2017. DOI: 10.48550/arXiv.1705.09783. arXiv: 1705.09783 [cs].
- [71] Paul R. Daugherty and H. James Wilson. *Human + Machine: Reimagining Work in the Age of AI*. Harvard Business Press, Mar. 2018. ISBN: 978-1-63369-387-6.
- [72] Peter Davis, Roy Lay-Yee, and Janet Pearson. "Using Micro-Simulation to Create a Synthesised Data Set and Test Policy Options: The Case of Health Service Effects under Demographic Ageing". In: *Health Policy (Amsterdam, Netherlands)* 97.2-3 (Oct. 2010), pp. 267–274. ISSN: 1872-6054. DOI: 10.1016/j.healthpol.2010.05.014.
- [73] Wesley Hanwen Deng et al. "Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM, June 2022, pp. 473–484. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533113.
- [74] Emily Denton et al. "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet". In: *Big Data & Society* 8.2 (July 2021), p. 20539517211035955. ISSN: 2053-9517. DOI: 10.1177/20539517211035955.
- [75] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. June 2021. DOI: 10.48550/arXiv.2105.05233. arXiv: 2105.05233 [cs, stat].
- [76] Mark Diaz et al. "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. June 2022, pp. 2342–2351. DOI: 10.1145/3531146.3534647. arXiv: 2206.08931 [cs].

- [77] Virginia Dignum. “RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES”. In: 1 (2017).
- [78] Paul Dourish. “Reading and Interpreting Ethnography”. In: *Ways of Knowing in HCI*. Ed. by Judith S. Olson and Wendy A. Kellogg. New York, NY: Springer New York, 2014, pp. 1–23. ISBN: 978-1-4939-0377-1 978-1-4939-0378-8. DOI: 10.1007/978-1-4939-0378-8_1.
- [79] Paul Dourish et al. “Reflective HCI: Towards a Critical Technical Practice”. In: *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*. Vienna Austria: ACM, Apr. 2004, pp. 1727–1728. ISBN: 978-1-58113-703-3. DOI: 10.1145/985921.986203.
- [80] Chris Dulhanty and Alexander Wong. *Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification*. Jan. 2020. arXiv: 2001.03071 [cs].
- [81] Cynthia Dwork. “Differential Privacy”. In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer, 2006, pp. 1–12. ISBN: 978-3-540-35908-1. DOI: 10.1007/11787006_1.
- [82] Cynthia Dwork and Deirdre K. Mulligan. “Privacy and Big Data: It’s Not Privacy, and It’s Not Fair”. In: *Stanford Law Review Online* 66 (2013), pp. 35–40.
- [83] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Journal of Privacy and Confidentiality* (2006). DOI: 10.29012/jpc.v7i3.405.
- [84] Upol Ehsan et al. “The Algorithmic Imprint”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. June 2022, pp. 1305–1317. DOI: 10.1145/3531146.3533186. arXiv: 2206.03275 [cs].
- [85] Emma Elder. “Wrongful Improvers as a Guiding Principle for Application of the FTC’s IP Deletion Requirement”. In: *WASHINGTON LAW REVIEW* 97.4 (2022).
- [86] Pardis Emami-Naeini et al. “Ask the Experts: What Should Be on an IoT Privacy and Security Label?” In: *2020 IEEE Symposium on Security and Privacy (SP)*. May 2020, pp. 447–464. DOI: 10.1109/SP40000.2020.00043.
- [87] *Engaging with the Public about Algorithmic Transparency in the Public Sector - Centre for Data Ethics and Innovation Blog*. June 2021.
- [88] Elena Esposito. “Algorithmic Memory and the Right to Be Forgotten on the Web”. In: *Big Data & Society* 4.1 (June 2017), p. 2053951717703996. ISSN: 2053-9517. DOI: 10.1177/2053951717703996.
- [89] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. *Quantifying the Influence of Context on Moral Judgments in Large Language Models*. 2022. arXiv: 2212.08073 [cs.CL].
- [90] Virginia Eubanks. *Automating Inequality*. St. Martin’s Press, 2018.
- [91] European Data Protection Board. *Guidelines 05/2020 on the Interplay of the Second Paragraph of Article 17 GDPR with Freedom of Expression and Information*. 2020.

- [92] European Parliament. *Artificial Intelligence Act (Provisional Agreement, 2024)*. 2024.
- [93] European Parliament. *General Data Protection Regulation*. 2018.
- [94] Hassan Ismail Fawaz et al. *Data Augmentation Using Synthetic Data for Time Series Classification with Deep Residual Networks*. Aug. 2018.
DOI: 10.48550/arXiv.1808.02455. arXiv: 1808.02455 [cs].
- [95] Federal Trade Commission. *Amazon Ring LLC—Complaint and Order*. 2023.
- [96] Federal Trade Commission. *Clearview AI, Inc.—Final Order*. 2022.
- [97] Federal Trade Commission. *In Re Everalbum, Inc. Complaint and Order*. 2021.
- [98] Federal Trade Commission. *In Re Rite Aid Corp. Proposed Decision and Order*. 2023.
- [99] D. Feldman. “Secrecy, Dignity, or Autonomy? Views of Privacy as a Civil Liberty”. In: *Current Legal Problems* 47.Part 2 (Jan. 1994), pp. 41–71. ISSN: 0070-1998, 2044-8422.
DOI: 10.1093/clp/47.Part_2.41.
- [100] Michael Feldman et al. *Certifying and Removing Disparate Impact*. July 2015.
DOI: 10.48550/arXiv.1412.3756. arXiv: 1412.3756 [cs, stat].
- [101] Thomas Palmeira Ferraz et al. *LLM Self-Correction with DeCRIM: Decompose, Critique, and Refine for Enhanced Following of Instructions with Multiple Constraints*. Oct. 2024.
- [102] “Fine-Grained Constraint Generation-Verification for Improved Instruction-Following”. In: *ACL Rolling Review - December 2024*. Feb. 2025.
- [103] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. “False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used across the Country to Predict Future Criminals. And It’s Biased against Blacks.”” In: *Federal Probation* 80.2 (2016), pp. 38–46.
- [104] Luciano Floridi. “Group Privacy: A Defence and an Interpretation”. In: *Philosophy & Technology* 27.1 (2014), pp. 1–3. DOI: 10.1007/978-3-319-46608-8_5.
- [105] UK Centre for Data Ethics et al. *BritainThinks: Complete Transparency, Complete Simplicity*. 2021.
- [106] Batya Friedman and David Hendry. “The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Austin Texas USA: ACM, May 2012, pp. 1145–1148. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2208562.
- [107] Chengguang Gan and Tatsunori Mori. “Sensitivity and Robustness of Large Language Models to Prompt Template in Japanese Text Classification Tasks”. In: *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 37)*. 2023, pp. 1–12.
- [108] Oscar H. Gandy. *The Panoptic Sort: A Political Economy of Personal Information*. Boulder, CO: Westview Press, 1993. ISBN: 9780813384551.

- [109] Matt Gardner et al. *Evaluating Models' Local Decision Boundaries via Contrast Sets*. Oct. 2020. DOI: 10.48550/arXiv.2004.02709. arXiv: 2004.02709 [cs].
- [110] Timnit Gebru et al. "Datasheets for Datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92. DOI: 10.1145/3458723.
- [111] Samuel Gehman et al. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. DOI: 10.18653/v1/2020.findings-emnlp.301.
- [112] *Genetic Information Nondiscrimination Act of 2008*. Public Law 110–233, 122 Stat. 881. 2008.
- [113] Michele E. Gilman. *Beyond Window Dressing: Public Participation for Marginalized Communities in the Datafied Society*. SSRN Scholarly Paper. Rochester, NY, Nov. 2022.
- [114] Anne N. Glucker et al. "Public Participation in Environmental Impact Assessment: Why, Who and How?" In: *Environmental Impact Assessment Review* 43 (Nov. 2013), pp. 104–111. ISSN: 0195-9255. DOI: 10.1016/j.eiar.2013.06.003.
- [115] Jake Goldenfein. "Privacy's Loose Grip on Facial Recognition: Law and the Operational Image". In: *The Cambridge Handbook of Facial Recognition in the Modern State*. Ed. by Monika Zalnieriute and Rita Matulionyte. Cambridge Law Handbooks. Cambridge: Cambridge University Press, 2024, pp. 74–86. ISBN: 978-1-009-32119-8. DOI: 10.1017/9781009321211.007.
- [116] Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. "Synthetic Data in Health Care: A Narrative Review". In: *PLOS Digital Health* 2.1 (Jan. 2023), e0000082. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000082.
- [117] Ian Goodfellow et al. "Generative Adversarial Nets". In: *NeurIPS 2014*. 2014, pp. 2672–2680. DOI: 10.48550/arXiv.1406.2661.
- [118] Bryce Goodman and Seth Flaxman. "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"". In: *AI Magazine* 38.3 (Oct. 2017), pp. 50–57. ISSN: 2371-9621. DOI: 10.1609/aimag.v38i3.2741.
- [119] Colin M. Gray et al. "Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–18. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445779.
- [120] Mary L. Gray and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, 2019.

- [121] Daniel Greene and Katie Shilton. “Platform Privacies: Governance, Collaboration, and the Different Meanings of “Privacy” in iOS and Android Development”. In: *New Media & Society* 20.4 (Apr. 2018), pp. 1640–1657. ISSN: 1461-4448. DOI: 10.1177/1461444817702397.
- [122] Glenn Greenwald. *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. 1. edition. New York: Metropolitan Books Henry Holt, 2014. ISBN: 978-1-62779-073-4.
- [123] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. Tech. rep. NIST IR 8280. National Institute of Standards and Technology (NIST), 2019. DOI: 10.6028/NIST.IR.8280.
- [124] Lara Groves et al. “Going Public: The Role of Public Participation Approaches in Commercial AI Labs”. In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. Chicago IL USA: ACM, June 2023, pp. 1162–1173. ISBN: 979-8-4007-0192-4. DOI: 10.1145/3593013.3594071.
- [125] Hangzhi Guo, Lizhen Zhu, and Ting-Hao Kenneth Huang. “Are GANs Biased? Evaluating GAN-Generated Facial Images via Crowdsourcing”. In: *Proceedings of the NeurIPS 2022 Workshop on Human Evaluation of Generative Models*. 2022.
- [126] Seda Gürses and Joris Van Hoboken. “Privacy after the Agile Turn”. In: (2016). DOI: 10.1017/9781316831960.032.
- [127] Kevin D. Haggerty and Richard V. Ericson. “The Surveillant Assemblage”. In: *British Journal of Sociology* 51.4 (2000), pp. 605–622. DOI: 10.4324/9781315243566-16.
- [128] John Halloran et al. “The Value of Values: Resourcing Co-Design of Ubiquitous Computing”. In: *CoDesign* 5.4 (Dec. 2009), pp. 245–273. ISSN: 1571-0882. DOI: 10.1080/15710880902920960.
- [129] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. “Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173582.
- [130] Donna Haraway. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective”. In: *Feminist Studies* 14.3 (1988), pp. 575–599. DOI: 10.4324/9780203427415-40.
- [131] Katie Harron et al. “Linking Data for Mothers and Babies in De-Identified Electronic Health Data”. In: *PLOS ONE* 11.10 (Oct. 2016), e0164667. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0164667.

- [132] Adam Harvey and Jules LaPlace. “Researchers Gone Wild: Origins and Endpoints of Image Training Datasets Created ”In the Wild””. In: *Design*. Ed. by Bianca Herlo et al. 1st ed. Vol. 54. Bielefeld, Germany: transcript Verlag, Dec. 2021, pp. 289–310. ISBN: 978-3-8376-5760-9 978-3-8394-5760-3. DOI: 10.14361/9783839457603-016.
- [133] Qianyu He et al. *Can Large Language Models Understand Real-World Complex Instructions?* Jan. 2024. DOI: 10.48550/arXiv.2309.09150. arXiv: 2309.09150 [cs].
- [134] *Hessisches Datenschutzgesetz (HDSG)*. Oct. 1970.
- [135] Anna Lauren Hoffmann. “Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse”. In: *Information, Communication & Society* 22.7 (2019), pp. 900–915. DOI: 10.1080/1369118X.2019.1573912.
- [136] Chengsong Hu et al. “Influence of Image Quality and Light Consistency on the Performance of Convolutional Neural Networks for Weed Mapping”. In: *Remote Sensing* 13.11 (Jan. 2021), p. 2140. ISSN: 2072-4292. DOI: 10.3390/rs13112140.
- [137] Hui Huang et al. *MuSC: Improving Complex Instruction Following with Multi-granularity Self-Contrastive Training*. Feb. 2025.
- [138] Stephan Hügel and Anna R. Davies. “Public Participation, Engagement, and Climate Change Adaptation: A Review of the Research Literature”. In: *Wiley Interdisciplinary Reviews. Climate Change* 11.4 (2020), e645. ISSN: 1757-7780. DOI: 10.1002/wcc.645.
- [139] Kirsty Hughes. “The Social Value of Privacy, the Value of Privacy to Society and Human Rights Discourse”. In: *Social Dimensions of Privacy: Interdisciplinary Perspectives*. Ed. by Beate Roessler and Dorota Mokrosinska. Cambridge: Cambridge University Press, 2015, pp. 225–243. ISBN: 978-1-107-05237-6. DOI: 10.1017/CB09781107280557.013.
- [140] Ben Hutchinson et al. “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 560–575. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445918.
- [141] Jevan Hutson and Ben Winters. “America’s next “Stop Model!”: Model Deletion”. In: *Georgetown Law Technology Review* 8.1 (2024), pp. 124–164.
- [142] Giovanni Iachello and Jason Hong. “End-User Privacy in Human-Computer Interaction”. In: *Foundations and Trends® in Human-Computer Interaction* 1.1 (2007), pp. 1–137. ISSN: 1551-3955, 1551-3963. DOI: 10.1561/11000000004.
- [143] Illinois General Assembly. *Biometric Information Privacy Act*. 2008.

- [144] Lilly Irani. *Chasing Innovation: Making Entrepreneurial Citizens in Modern India*. Princeton Studies in Culture and Technology. Princeton, New Jersey ; Oxford, United Kingdom: Princeton University Press, 2019. ISBN: 978-0-691-17514-0 978-0-691-17513-3.
- [145] Golnaz Arastoopour Irgens et al. “Designing with and for Youth: A Participatory Design Research Approach for Critical Machine Learning Education”. In: *Educational Technology & Society* 25.4 (2022), pp. 126–141. ISSN: 1176-3647. JSTOR: 48695986.
- [146] Ole Sejer Iversen, Kim Halskov, and Tuck W. Leong. “Values-Led Participatory Design”. In: *CoDesign* 8.2-3 (June 2012), pp. 87–103. ISSN: 1571-0882. DOI: 10.1080/15710882.2012.672575.
- [147] Steven J. Jackson. “Rethinking Repair”. In: *Media Technologies*. Ed. by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. The MIT Press, Feb. 2014, pp. 221–240. ISBN: 978-0-262-52537-4. DOI: 10.7551/mitpress/9780262525374.003.0011.
- [148] Benjamin N Jacobsen. “Machine Learning and the Politics of Synthetic Data”. In: *Big Data & Society* 10.1 (Jan. 2023), p. 20539517221145372. ISSN: 2053-9517. DOI: 10.1177/20539517221145372.
- [149] Benjamin N Jacobsen. “Machine Learning and the Politics of Synthetic Data”. In: *Big Data & Society* 10.1 (Jan. 2023), p. 20539517221145372. ISSN: 2053-9517. DOI: 10.1177/20539517221145372.
- [150] Yuxin Jiang et al. *FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models*. June 2024. DOI: 10.48550/arXiv.2310.20410. arXiv: 2310.20410 [cs].
- [151] Adam Joinson et al. “Privacy, Trust, and Self-Disclosure Online”. In: *Human-Computer Interaction* 25.1 (Jan. 2010), pp. 1–24. ISSN: 0737-0024. DOI: 10.1080/07370020903586662.
- [152] James Jordon et al. *Synthetic Data – What, Why and How?* May 2022. arXiv: 2205.03257 [cs].
- [153] Faisal Kamiran and Toon Calders. “Classifying without Discriminating”. In: *Control and Communication 2009 2nd International Conference on Computer*. Feb. 2009, pp. 1–6. DOI: 10.1109/IC4.2009.4909197.
- [154] Davinder Kaur et al. “Trustworthy Artificial Intelligence: A Review”. In: *ACM Computing Surveys* 55.2 (Feb. 2023), pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3491209.
- [155] Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, eds. *Advances in Face Detection and Facial Image Analysis*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-25956-7 978-3-319-25958-1. DOI: 10.1007/978-3-319-25958-1.

- [156] Mehtab Khan and Alex Hanna. *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*. SSRN Scholarly Paper. Rochester, NY, Sept. 2022. DOI: 10.2139/ssrn.4217148.
- [157] Jennifer L. King. “Privacy, Trust, and Design: A Digital Systems Perspective”. PhD thesis. Berkeley, CA: University of California, Berkeley, 2018.
- [158] John Kirchenbauer et al. “A Watermark for Large Language Models”. In: *arXiv preprint arXiv:2301.10226* (2023). arXiv: 2301.10226.
- [159] Bran Knowles et al. “A Scenario-Based Methodology for Exploring Risks: Children and Programmable IoT”. In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. DIS ’19. New York, NY, USA: Association for Computing Machinery, June 2019, pp. 751–761. ISBN: 978-1-4503-5850-7. DOI: 10.1145/3322276.3322315.
- [160] Nitin Kohli and Joshua E. Blumenstock. “Enabling Humanitarian Applications with Targeted Differential Privacy”. In: *arXiv preprint arXiv:2408.13424* (2024). arXiv: 2408.13424.
- [161] Adam Kortylewski et al. “Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Long Beach, CA, USA: IEEE, June 2019, pp. 2261–2268. ISBN: 978-1-7281-2506-0. DOI: 10.1109/CVPRW.2019.00279.
- [162] Yuri Kuratov et al. *BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack*. Nov. 2024. DOI: 10.48550/arXiv.2406.10149. arXiv: 2406.10149 [cs].
- [163] Markus Langer and Cornelius J. König. “Introducing and Testing the Creepiness of Situation Scale (CRoSS)”. In: *Frontiers in Psychology* 9 (2018). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.02220.
- [164] Nicol Turner Lee, Paul Resnick, and Genie Barton. *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*. May 2019.
- [165] Haoran Li et al. *PrivaCI-Bench: Evaluating Privacy with Contextual Integrity and Legal Compliance*. Feb. 2025. DOI: 10.48550/arXiv.2502.17041. arXiv: 2502.17041 [cs].
- [166] Oliver Li et al. “NormDial: A Comparable Bilingual Synthetic Dialog Dataset for Modeling Social Norm Adherence and Violation”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15732–15744. DOI: 10.18653/v1/2023.emnlp-main.974.
- [167] Tiffany C. Li. *Algorithmic Destruction*. SSRN Scholarly Paper. Rochester, NY, 2022. DOI: 10.2139/ssrn.4066845.

- [168] Yuqing Liang et al. *Watermarking Techniques for Large Language Models: A Survey*. Aug. 2024. DOI: 10.48550/arXiv.2409.00089. arXiv: 2409.00089 [cs].
- [169] Q. Vera Liao, Daniel Gruen, and Sarah Miller. “Questioning the AI: Informing Design Practices for Explainable AI User Experiences”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–15. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376590.
- [170] Gili Lior et al. *WildIFEval: Instruction Following in the Wild*. Mar. 2025. DOI: 10.48550/arXiv.2503.06573. arXiv: 2503.06573 [cs].
- [171] Jialin Liu et al. “Deep Learning for Procedural Content Generation”. In: *Neural Computing and Applications* 33.1 (Jan. 2021), pp. 19–37. ISSN: 1433-3058. DOI: 10.1007/s00521-020-05383-8.
- [172] Ken Liu. *Machine Unlearning in 2024*.
- [173] Xiang Liu et al. “LongGenBench: Long-context Generation Benchmark”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 865–883. DOI: 10.18653/v1/2024.findings-emnlp.48.
- [174] Andrei Lopatenko. *Alopatenko/LLMEvaluation*. Apr. 2025.
- [175] Edgar Alonso Lopez-Rojas and Stefan Axelsson. “Money Laundering Detection Using Synthetic Data”. In: May 2012.
- [176] Aengus Lynch et al. *Eight Methods to Evaluate Robust Unlearning in LLMs*. Feb. 2024. DOI: 10.48550/arXiv.2402.16835. arXiv: 2402.16835 [cs].
- [177] David Lyon. *Surveillance Studies: An Overview*. Cambridge: Polity Press, 2007. ISBN: 978-0-7456-3593-4.
- [178] Kirsten Martin. “The Penalty for Privacy Violations: How Privacy Violations Impact Trust Online”. In: *Journal of Business Research* 82 (Jan. 2018), pp. 103–116. ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2017.08.034.
- [179] Kirsten Martin and Katie Shilton. “Putting Mobile Application Privacy in Context: An Empirical Study of User Privacy Expectations for Mobile Devices”. In: *The Information Society* 32.3 (May 2016), pp. 200–216. ISSN: 0197-2243, 1087-6537. DOI: 10.1080/01972243.2016.1153012.
- [180] Kirsten Martin and Katie Shilton. “Why Experience Matters to Privacy: How Context-based Experience Moderates Consumer Privacy Expectations for Mobile Applications”. In: *Journal of the Association for Information Science and Technology* 67.8 (Aug. 2016), pp. 1871–1882. ISSN: 2330-1635, 2330-1643. DOI: 10.1002/asi.23500.

- [181] Gonzalo Martínez et al. *Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet*. June 2023. DOI: 10.48550/arXiv.2306.06130. arXiv: 2306.06130 [cs].
- [182] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. “Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–23. ISSN: 2573-0142. DOI: 10.1145/3359174.
- [183] Sean McGregor. “Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.17 (May 2021), pp. 15458–15463. ISSN: 2374-3468. DOI: 10.1609/aaai.v35i17.17817.
- [184] Yifei Ming et al. *FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon Is Made of Marshmallows"*. Oct. 2024. DOI: 10.48550/arXiv.2410.03727. arXiv: 2410.03727 [cs].
- [185] Ilya Mironov. “Rényi Differential Privacy”. In: *IEEE Transactions on Information Theory* 65.5 (2019), pp. 3025–3048. DOI: 10.1109/csf.2017.11.
- [186] Yutao Mou, Shikun Zhang, and Wei Ye. *SG-Bench: Evaluating LLM Safety Generalization Across Diverse Tasks and Prompt Types*. Oct. 2024. DOI: 10.48550/arXiv.2410.21965. arXiv: 2410.21965 [cs].
- [187] Deirdre K. Mulligan and Kenneth A. Bamberger. “Saving Governance-By-Design”. In: *California Law Review* 106.3 (2018), pp. 697–784. ISSN: 0008-1221. JSTOR: 26577731.
- [188] Deirdre K. Mulligan and Nick Doty. *Design Wars: The FBI, Apple and Hundreds of Millions of Phones*. 2016.
- [189] Deirdre K. Mulligan, Colin Koopman, and Nick Doty. “Privacy Is an Essentially Contested Concept: A Multi-Dimensional Analytic for Mapping Privacy”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083 (Dec. 2016), p. 20160118. DOI: 10.1098/rsta.2016.0118.
- [190] Ambar Murillo et al. ““If I Press Delete, It’s Gone” - User Understanding of Online Data Deletion and Expiration”. In: *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. 2018, pp. 329–339. ISBN: 978-1-939133-10-6.
- [191] Rudra Murthy et al. *Evaluating the Instruction-following Abilities of Language Models Using Knowledge Tasks*. Oct. 2024. DOI: 10.48550/arXiv.2410.12972. arXiv: 2410.12972 [cs].
- [192] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. “Descent-to-Delete: Gradient-Based Methods for Machine Unlearning”. In: *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. PMLR, Mar. 2021, pp. 931–962.

- [193] Che Ngufor et al. “Mixed Effect Machine Learning: A Framework for Predicting Longitudinal Change in Hemoglobin A1c”. In: *Journal of Biomedical Informatics* 89 (Jan. 2019), pp. 56–67. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2018.09.001.
- [194] Xuan-Phi Nguyen et al. *SFR-RAG: Towards Contextually Faithful LLMs*. Sept. 2024. DOI: 10.48550/arXiv.2409.09916. arXiv: 2409.09916 [cs].
- [195] Sergey I. Nikolenko. *Synthesis Humans. Create Perfectly Labeled Images & Video for ML Models*. Nov. 2022.
- [196] Helen Nissenbaum. “Privacy as Contextual Integrity”. In: *Washington Law Review* 79.1 (2004), pp. 119–158.
- [197] Safiya Umoja Noble. *Algorithms of Oppression*. New York University Press, 2018.
- [198] Justin Norman, Shruti Agarwal, and Hany Farid. *An Evaluation of Forensic Facial Recognition*. Nov. 2023. DOI: 10.48550/arXiv.2311.06145. arXiv: 2311.06145 [cs].
- [199] Organisation for Economic Co-operation and Development. “Guidelines on the Protection of Privacy and Transborder Flows of Personal Data”. In: (1980). DOI: 10.1787/9789264196391-en.
- [200] Long Ouyang et al. *Training Language Models to Follow Instructions with Human Feedback*. Mar. 2022. DOI: 10.48550/arXiv.2203.02155. arXiv: 2203.02155 [cs].
- [201] Junwon Park et al. “AI-Based Request Augmentation to Increase Crowdsourcing Participation”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), pp. 115–124. ISSN: 2769-1349. DOI: 10.1609/hcomp.v7i1.5282.
- [202] Mihir Parmar et al. *LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models*. June 2024. DOI: 10.48550/arXiv.2404.15522. arXiv: 2404.15522 [cs].
- [203] Frank Pasquale. *Licensure as Data Governance*. Oct. 2021.
- [204] Reema Patel et al. *Participatory Data Stewardship*. Tech. rep. Ada Lovelace Institute, 2021.
- [205] Amandalynne Paullada et al. “Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research”. In: *Patterns* 2.11 (Nov. 2021), p. 100336. ISSN: 2666-3899. DOI: 10.1016/j.patter.2021.100336.
- [206] Martin Pawelczyk et al. *Machine Unlearning Fails to Remove Data Poisoning Attacks*. June 2024. DOI: 10.48550/arXiv.2406.17216. arXiv: 2406.17216 [cs].
- [207] Martin Pawelczyk et al. *On the Trade-Off between Actionable Explanations and the Right to Be Forgotten*. Oct. 2022. DOI: 10.48550/arXiv.2208.14137. arXiv: 2208.14137 [cs].

- [208] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. *Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers*. Nov. 2021. DOI: 10.48550/arXiv.2108.02922. arXiv: 2108.02922 [cs].
- [209] Daniel Perez. *ChatGPT: Learning Tool — or Threat? How a Texas College Is Eyeing New AI Program*. Jan. 2023.
- [210] Alina Pommeranz et al. “Elicitation of Situated Values: Need for Tools to Help Stakeholders and Designers to Reflect and Communicate”. In: *Ethics and Information Technology* 14.4 (Dec. 2012), pp. 285–303. ISSN: 1572-8439. DOI: 10.1007/s10676-011-9282-6.
- [211] William L. Prosser. “Privacy”. In: *California Law Review* 48.3 (1960), pp. 383–423. DOI: 10.2307/3478805.
- [212] Ismini Psychoula, Liming Chen, and Feng Chen. “Privacy Modelling and Management for Assisted Living within Smart Homes”. In: *2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom)*. Oct. 2017, pp. 1–6. DOI: 10.1109/HealthCom.2017.8210782.
- [213] Haibo Qiu et al. “SynFace: Face Recognition With Synthetic Data”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10880–10890. DOI: 10.1109/iccv48922.2021.01070.
- [214] Haoyi Qiu et al. *Evaluating Cultural and Social Awareness of LLM Web Agents*. Mar. 2025. DOI: 10.48550/arXiv.2410.23252. arXiv: 2410.23252 [cs].
- [215] Rafael Rafailov et al. *Direct Preference Optimization: Your Language Model Is Secretly a Reward Model*. July 2024. DOI: 10.48550/arXiv.2305.18290. arXiv: 2305.18290 [cs].
- [216] Manish Raghavan et al. “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* ’20*. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 469–481. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372828.
- [217] Stephanie Carroll Rainie et al. *Indigenous Data Sovereignty*. African Minds and the International Development Research Centre (IDRC), 2019, pp. 300–319. ISBN: 978-1-928331-95-7.
- [218] Inioluwa Deborah Raji and Genevieve Fried. *About Face: A Survey of Facial Recognition Evaluation*. Feb. 2021. DOI: 10.48550/arXiv.2102.00813. arXiv: 2102.00813 [cs].
- [219] Inioluwa Deborah Raji et al. *Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing*. Jan. 2020. DOI: 10.48550/arXiv.2001.00964. arXiv: 2001.00964 [cs].

- [220] Abhinav Rao et al. *NormAd: A Benchmark for Measuring the Cultural Adaptability of Large Language Models*. July 2024. DOI: 10.48550/arXiv.2404.12464. arXiv: 2404.12464 [cs].
- [221] Yulia Razmetaeva. “The Right to Be Forgotten in the European Perspective”. In: *TalTech Journal of European Studies* 10.1 (May 2020), pp. 58–76. DOI: 10.1515/bjes-2020-0004.
- [222] Benjamin Recht et al. *Do ImageNet Classifiers Generalize to ImageNet?* June 2019. DOI: 10.48550/arXiv.1902.10811. arXiv: 1902.10811 [cs, stat].
- [223] Priscilla M. Regan. *Legislating Privacy: Technology, Social Values, and Public Policy*. University of North Carolina Press, 1995.
- [224] Samantha Robertson and Niloufar Salehi. *What If I Don’t Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design*. July 2020. DOI: 10.48550/arXiv.2007.06718. arXiv: 2007.06718 [cs].
- [225] Isaac Robinson and John Burden. *Framing the Game: How Context Shapes LLM Decision-Making*. Mar. 2025. DOI: 10.48550/arXiv.2503.04840. arXiv: 2503.04840 [cs].
- [226] Mohammad Al-Rubaie and J. Morris Chang. “Privacy-Preserving Machine Learning: Threats and Solutions”. In: *IEEE Security & Privacy* 17.2 (Mar. 2019), pp. 49–58. ISSN: 1558-4046. DOI: 10.1109/MSEC.2018.2888775.
- [227] Soumya Sanyal, Harman Singh, and Xiang Ren. “FaiRR: Faithful and Robust Deductive Reasoning over Natural Language”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1075–1093. DOI: 10.18653/v1/2022.acl-long.77.
- [228] Morgan Klaus Scheuerman et al. “From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023), 55:1–55:33. DOI: 10.1145/3579488.
- [229] Morgan Klaus Scheuerman et al. “How We’ve Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (May 2020), 58:1–58:35. DOI: 10.1145/3392866.
- [230] Eva-Maria Schomakers, Chantal Lidynia, and Martina Zieffle. “All of Me? Users’ Preferences for Privacy-Preserving Data Markets and the Importance of Anonymity”. In: *Electronic Markets* 30.3 (Sept. 2020), pp. 649–665. ISSN: 1422-8890. DOI: 10.1007/s12525-020-00404-9.
- [231] Paul Schwartz. “Global Data Privacy: The EU Way Symposium”. In: *New York University Law Review* 94.4 (2019), pp. 771–818.

- [232] Ayush Sekhari et al. *Remember What You Want to Forget: Algorithms for Machine Unlearning*. July 2021. DOI: 10.48550/arXiv.2103.03279. arXiv: 2103.03279 [cs].
- [233] Andrew D. Selbst et al. “Fairness and Abstraction in Sociotechnical Systems”. In: *Fat* 2019*. 2019, pp. 59–68. DOI: 10.1145/3287560.3287598.
- [234] Daniela Sele and Marina Chugunova. *Putting a Human in the Loop: Increasing Uptake, but Decreasing Accuracy of Automated Decision-Making*. SSRN Scholarly Paper. Rochester, NY, Nov. 2022.
- [235] Norbert Seyff et al. “Mobile Discovery of Requirements for Context-Aware Systems”. In: *Requirements Engineering: Foundation for Software Quality*. Ed. by Barbara Paech and Colette Rolland. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 183–197. ISBN: 978-3-540-69062-7. DOI: 10.1007/978-3-540-69062-7_18.
- [236] Norbert Seyff et al. “Scenarios in the Wild: Experiences with a Contextual Requirements Discovery Method”. In: *Requirements Engineering: Foundation for Software Quality*. Ed. by Martin Glinz and Patrick Heymans. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 147–161. ISBN: 978-3-642-02050-6. DOI: 10.1007/978-3-642-02050-6_13.
- [237] Renee Shelby et al. “Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction”. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 723–741. ISBN: 979-8-4007-0231-0. DOI: 10.1145/3600211.3604673.
- [238] Donghee Shin. “The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI”. In: *International Journal of Human-Computer Studies* 146 (Feb. 2021), p. 102551. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2020.102551.
- [239] Ilia Shumailov et al. *The Curse of Recursion: Training on Generated Data Makes Models Forget*. May 2023. DOI: 10.48550/arXiv.2305.17493. arXiv: 2305.17493 [cs].
- [240] Yan Shvartzshnaider and Vasisht Duddu. *Position: Contextual Integrity Washing for Language Models*. Jan. 2025.
- [241] Seyed Hossein Siadat and Minseok Song. “Understanding Requirement Engineering for Context-Aware Service-Based Applications”. In: 2012 (Aug. 2012). DOI: 10.4236/jsea.2012.58062.
- [242] Mona Sloane. “To Make AI Fair, Here’s What We Must Learn to Do”. In: *Nature* 605 (May 2022), pp. 9–9. DOI: 10.1038/d41586-022-01202-3.
- [243] Mona Sloane et al. *Participation Is Not a Design Fix for Machine Learning*. Aug. 2020. DOI: 10.48550/arXiv.2007.02423. arXiv: 2007.02423 [cs].

- [244] Nathalie A. Smuha. “From a ‘Race to AI’ to a ‘Race to AI Regulation’: Regulatory Competition for Artificial Intelligence”. In: *Law, Innovation and Technology* 13.1 (Jan. 2021), pp. 57–84. ISSN: 1757-9961. DOI: 10.1080/17579961.2021.1898300.
- [245] Robert Soden, Austin Toombs, and Michaelanne Thomas. “Evaluating Interpretive Research in HCI”. In: *Interactions* 31.1 (Jan. 2024), pp. 38–42. ISSN: 1072-5520, 1558-3449. DOI: 10.1145/3633200.
- [246] Daniel J. Solove. *Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data*. SSRN Scholarly Paper. Rochester, NY, Jan. 2024. DOI: 10.2139/ssrn.4322198.
- [247] Daniel J. Solove. “Privacy Self-management and the Consent Dilemma”. In: *Harvard Law Review* 126.7 (2013), pp. 1880–1903.
- [248] Daniel J. Solove and Woodrow Hartzog. “The FTC and the New Common Law of Privacy”. In: *Columbia Law Review* 114.3 (2014), pp. 583–676.
- [249] Dean Spade. *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law*. Duke University Press, July 2015. ISBN: 978-0-8223-7479-4.
- [250] A. et al. Srivastava. *Beyond the Imitation Game Benchmark (BIG-bench)*. 2023.
- [251] Meg Strickler. “Recent Developments in Privacy Law”. In: *Business Lawyer* 76 (2021), p. 269.
- [252] Guangzhi Sun et al. *CASE-Bench: Context-Aware SafeTy Benchmark for Large Language Models*. Feb. 2025. DOI: 10.48550/arXiv.2501.14940. arXiv: 2501.14940 [cs].
- [253] Wangtao Sun et al. “Beyond Instruction Following: Evaluating Inferential Rule Following of Large Language Models”. In: (2024). DOI: 10.48550/ARXIV.2407.08440.
- [254] Julie Taboh. *Facial Recognition Technology Solves Crimes, but at What Cost?* June 2021.
- [255] Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. *ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language*. June 2021. DOI: 10.48550/arXiv.2012.13048. arXiv: 2012.13048 [cs].
- [256] Alex S. Taylor et al. “Data-in-Place: Thinking through the Relations Between Data and Community”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul Republic of Korea: ACM, Apr. 2015, pp. 2863–2872. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702558.
- [257] David Thiel. *Identifying and Eliminating CSAM in Generative ML Training Data and Models*. Tech. rep. Stanford Internet Observatory, Dec. 2023.
- [258] Anvith Thudi et al. “On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning”. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 4007–4022. ISBN: 978-1-939133-31-1.

- [259] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. July 2023. DOI: 10.48550/arXiv.2307.09288. arXiv: 2307.09288 [cs].
- [260] Trieu H. Trinh et al. “Solving Olympiad Geometry without Human Demonstrations”. In: *Nature* 625.7995 (Jan. 2024), pp. 476–482. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06747-5.
- [261] Joseph Turow and Michael Hennessy. “Internet Privacy and Institutional Trust: Insights from a National Survey”. In: *New Media & Society* 9.2 (Apr. 2007), pp. 300–318. ISSN: 1461-4448. DOI: 10.1177/1461444807072219.
- [262] U. S. Department of Health, Education, and Welfare. *Records, Computers and the Rights of Citizens*. 1973.
- [263] Salomé Viljoen. “A Relational Theory of Data Governance”. In: *The Yale Law Journal* 131.2 (2021), pp. 573–654.
- [264] Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. “Humans Forget, Machines Remember: Artificial Intelligence and the Right to Be Forgotten”. In: *Computer Law & Security Review* 34.2 (Apr. 2018), pp. 304–313. ISSN: 0267-3649. DOI: 10.1016/j.clsr.2017.08.007.
- [265] Jonathan Vitale et al. “Be More Transparent and Users Will Like You: A Robot Privacy and User Experience Design Experiment”. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’18. New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 379–387. ISBN: 978-1-4503-4953-6. DOI: 10.1145/3171221.3171269.
- [266] W. Gregory Voss. *The CCPA and the GDPR Are Not the Same: Why You Should Understand Both*. SSRN Scholarly Paper. Rochester, NY, Jan. 2021.
- [267] Minzheng Wang et al. *Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA*. Oct. 2024. DOI: 10.48550/arXiv.2406.17419. arXiv: 2406.17419 [cs].
- [268] Xiaoqian Wang and Heng Huang. *Approaching Machine Learning Fairness through Adversarial Network*. Sept. 2019. DOI: 10.48550/arXiv.1909.03013. arXiv: 1909.03013 [cs, stat].
- [269] Zhaoyang Wang et al. *Verifiable Format Control for Large Language Model Generations*. Feb. 2025. DOI: 10.48550/arXiv.2502.04498. arXiv: 2502.04498 [cs].
- [270] Samuel D. Warren and Louis D. Brandeis. “The Right to Privacy”. In: *Harvard Law Review* 4.5 (1890), pp. 193–220. DOI: 10.2307/1321160.
- [271] Samuel Wehrli et al. “Bias, Awareness, and Ignorance in Deep-Learning-Based Face Recognition”. In: *AI and Ethics* 2.3 (Aug. 2022), pp. 509–522. ISSN: 2730-5961. DOI: 10.1007/s43681-021-00108-6.
- [272] Laura Weidinger et al. “Toward an Evaluation Science for Generative AI Systems”. In: *arXiv preprint arXiv:2503.05336* (2025). arXiv: 2503.05336.

- [273] Bosi Wen et al. *Benchmarking Complex Instruction-Following with Multiple Constraints Composition*. Oct. 2024. DOI: 10.48550/arXiv.2407.03978. arXiv: 2407.03978 [cs].
- [274] Alan F. Westin. *Privacy and Freedom*. Atheneum, 1967.
- [275] Cedric Deslandes Whitney et al. “HCI Tactics for Politics from Below: Meeting the Challenges of Smart Cities”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, May 2021, pp. 1–15. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445314.
- [276] Maranke Wieringa. “What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 1–18. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372833.
- [277] Richmond Y. Wong et al. “Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks”. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (Dec. 2017), 111:1–111:26. DOI: 10.1145/3134746.
- [278] Sebastien C. Wong et al. *Understanding Data Augmentation for Classification: When to Warp?* Nov. 2016. arXiv: 1609.08764 [cs].
- [279] Erroll Wood et al. “Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3681–3691. DOI: 10.1109/iccv48922.2021.00366.
- [280] Yuhao Wu et al. *LongGenbench: Benchmarking Long-Form Generation in Long Context LLMs*. Sept. 2024. DOI: 10.48550/arXiv.2409.02076. arXiv: 2409.02076 [cs].
- [281] Zhiyan Wu et al. “The Impact of Algorithmic Price Discrimination on Consumers’ Perceived Betrayal”. In: *Frontiers in Psychology* 13 (2022). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.825420.
- [282] Susan Wyche. “The Benefits of Using Design Workbooks with Speculative Design Proposals in Information Communication Technology for Development (ICTD)”. In: *Designing Interactive Systems Conference 2021* (June 2021), pp. 1861–1874. DOI: 10.1145/3461778.3462140.
- [283] Congying Xia et al. *FOFO: A Benchmark to Evaluate LLMs’ Format-Following Capability*. Feb. 2024. DOI: 10.48550/arXiv.2402.18667. arXiv: 2402.18667 [cs].
- [284] Austin Xu et al. *Does Context Matter? ContextualJudgeBench for Evaluating LLM-based Judges in Contextual Settings*. Mar. 2025. DOI: 10.48550/arXiv.2503.15620. arXiv: 2503.15620 [cs].

- [285] Kaiyu Yang et al. “Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Jan. 2020, pp. 547–558. DOI: 10.1145/3351095.3375709. arXiv: 1912.07726 [cs].
- [286] Yaxing Yao et al. “Defending My Castle: A Co-Design Study of Privacy Mechanisms for Smart Homes”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, May 2019, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300428.
- [287] Rui-Jie Yew and Alice Xiang. “Regulating Facial Processing Technologies: Tensions Between Legal and Technical Considerations in the Application of Illinois BIPA”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM, June 2022, pp. 1017–1027. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533163.
- [288] Dong Yi et al. *Learning Face Representation from Scratch*. Nov. 2014. arXiv: 1411.7923 [cs].
- [289] Lu Zhang, Yongkai Wu, and Xintao Wu. *A Causal Framework for Discovering and Removing Direct and Indirect Discrimination*. Nov. 2016. DOI: 10.48550/arXiv.1611.07509. arXiv: 1611.07509 [cs].
- [290] Tao Zhang et al. *CFBench: A Comprehensive Constraints-Following Benchmark for LLMs*. Aug. 2024. DOI: 10.48550/arXiv.2408.01122. arXiv: 2408.01122 [cs].
- [291] Zhexin Zhang et al. *Agent-SafetyBench: Evaluating the Safety of LLM Agents*. Dec. 2024. DOI: 10.48550/arXiv.2412.14470. arXiv: 2412.14470 [cs].
- [292] Jeffrey Zhou et al. *Instruction-Following Evaluation for Large Language Models*. Nov. 2023.
- [293] Andy Zou et al. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. Dec. 2023. DOI: 10.48550/arXiv.2307.15043. arXiv: 2307.15043 [cs].
- [294] Shoshana Zuboff. *The Age of Surveillance Capitalism*. PublicAffairs, 2019.